

Indice

Introduzione	1
1 Tecnologie CMOS bulk e MOSFET a body ultra sottile	5
1.1 Scaling delle tecnologie CMOS	5
1.1.1 Teoria dello scaling a campo costante	8
1.1.2 Teoria generalizzata dello scaling	14
1.2 Limiti nello scaling delle tecnologie CMOS bulk planari	16
1.2.1 Limiti fisici: tunneling nell'ossido di gate ed effetti di canale corto	19
1.2.2 Limiti tecnologici e dei materiali	27
1.3 Tecnologie MOSFET a body ultra sottile	29
1.3.1 Tecnologie fully depleted SOI	31
1.3.2 Tecnologie CMOS a gate multiplo	36
1.3.3 Tecnologie FinFET	39
2 Caratterizzazione statica in dispositivi FinFET	45
2.1 Descrizione dei dispositivi sotto misura	45
2.2 Caratteristiche statiche	48
2.2.1 Corrente di drain caratteristica normalizzata	61
2.2.2 Guadagno di tensione intrinseco	66
2.2.3 Corrente di perdita di gate	70
2.2.4 Estrazione della tensione di soglia	70
2.2.5 Pendenza della caratteristica $I_D - V_{GS}$ in sottosoglia	76
3 Caratterizzazione di rumore in dispositivi FinFET	81
3.1 Sorgenti equivalenti di rumore nei circuiti lineari	81
3.1.1 Parametri e modelli di rumore per dispositivi CMOS nanometrici	83
3.2 Strumentazione per misure di rumore	89

3.2.1	Setup di misura	90
3.2.2	Amplificatore a transimpedenza e rete di polarizzazione	93
3.2.3	Analisi delle sorgenti di rumore	99
3.3	Misure di rumore	104
3.3.1	Misure di densità spettrale di rumore	104
3.3.2	Dipendenza del rumore dalle dimensioni del gate	109
3.3.3	Analisi dei parametri di rumore	110
Conclusioni		115
Bibliografia		117

Introduzione

La rapidissima espansione della porzione del mercato basata sulle tecnologie microelettroniche (computer, telecomunicazioni, strumentazione, anche biomedica, intrattenimento, automotive, per citare alcuni settori) ha fornito e continua a fornire un notevole slancio all'evoluzione di tali tecnologie. Nel caso specifico, per evoluzione si intende la spinta verso la miniaturizzazione e la riduzione della potenza dissipata (ma anche verso un incremento della complessità dei processi produttivi, per esempio con la possibilità di integrare in una medesima tecnologia dispositivi a bassa potenza per funzioni logiche con dispositivi ad alta tensione) finalizzata ad un aumento della densità di funzioni, ad un incremento della velocità operativa e ad un maggior ritorno economico.

Le tecnologie *Complementary Metal-Oxide Semiconductor* (CMOS) bulk planari hanno però ormai raggiunto un limite (28 - 22 nm) oltre il quale ogni ulteriore riduzione delle dimensioni comporterebbe una serie di problemi che semplici miglioramenti del processo tecnologico preesistente non riuscirebbero a risolvere. Questi accorgimenti, quali ad esempio *halo implantation*, correzione ottica nei processi litografici, uso del nitruro insieme all'ossido di silicio negli strati isolanti di gate, ossidi a bassa costante dielettrica utilizzati come isolanti intermetallici, hanno tuttavia permesso di superare ostacoli che apparivano invalicabili, così che la previsione di Moore circa l'evoluzione dei processi microelettronici non è stata smentita, ameno fino ad oggi. I problemi cui si fa riferimento sono di diversa natura e riguardano:

- la densità della potenza dissipata;
- la riduzione delle dimensioni (*tunneling*, effetti di canale corto);
- i materiali (isolamento e conduzione);
- i limiti dei processi litografici.

In modo particolare l'aumento della potenza dissipata per unità di superficie trae origine dal fatto che il droggaggio di canale e le tensioni di soglia e di

alimentazione non scalano così facilmente come le dimensioni geometriche. È pertanto necessario ridurre lo spessore dell'ossido di gate per limitare gli effetti del punch-through tra source e drain, ridurre il fenomeno del Drain Induced Barrier Lowering (DIBL) e più in generale mitigare gli effetti di canale corto. Questo tuttavia comporta un aumento intollerabile della corrente di gate, che scala esponenzialmente con la riduzione dell'ossido.

In particolare, le due tecnologie che si stanno proponendo all'attenzione degli operatori del settore e che appaiono tra le più promettenti in vista del proseguimento del processo di scaling tecnologico sono:

- la tecnologia Fully-Depleted Silicon-On-Insulator CMOS (FD SOI CMOS);
- la tecnologia multiple gate CMOS; queste includono le tecnologie *Double e Triple gate CMOS* (DG e TG CMOS), delle quali i FinFET rappresentano un caso particolare.

Con il ricorso a questi processi innovativi si prevede di arrivare a minime lunghezze di canale pari a 7 nm entro il 2020, con prestazioni più che soddisfacenti in termini, per esempio, di dissipazione di potenza e rapporto Ion/Ioff. L'idea è quella di massimizzare il controllo del terminale di gate sul canale e minimizzare le variazioni statistiche (si pensi ad esempio alla variazione della tensione di soglia) che possono pregiudicare le prestazioni del dispositivo.

Le due soluzioni proposte sfruttano entrambe un *ultra-thin body* (la regione nella quale si realizza il canale del dispositivo) estremamente sottile, ma dal punto di vista progettuale risultano estremamente differenti. Il FD-SOI MOSFET è realizzato con un sottile strato di silicio che viene deposto al di sopra di uno strato isolante, il *buried oxide* (BOX). In un dispositivo SOI gli effetti di canale corto, come vedremo nel seguito, sono controllati dallo spessore del body di silicio, mentre la variazione della tensione di soglia dovuta all'effetto di *Random Dopant Fluctuation* (RDF) è sensibilmente ridotta mediante la realizzazione di un canale non drogato. In un FinFET la regione di canale sporge dalla superficie di silicio, dando origine ad una pinna (*fin*) che costituisce il body del dispositivo. Il gate ricopre tre lati della pinna (il quarto coincide con la superficie del wafer di silicio), così da ottenere un miglior controllo degli effetti di canale corto. Questi dipendono soprattutto dallo spessore della pinna, il quale diventa un parametro critico per la realizzazione del dispositivo.

Trattandosi di tecnologie innovative, la loro caratterizzazione sperimentale riveste un interesse particolare. Infatti, la caratterizzazione in termini di rumore dei dispositivi a semiconduttore è di primaria importanza al fine di comprendere i limiti di un dato processo tecnologico. Per analizzare le prestazioni di

rumore di un circuito è necessario disporre di modelli che permettano la descrizione delle sorgenti di rumore nei componenti circuitali elementari, quali, per esempio, resistori, diodi e transistori ad effetto di campo. I principali modelli a cui si fa riferimento nella trattazione comprendono: rumore termico, rumore con spettro di tipo 1/f (o rumore *flicker*) e rumore di tipo Lorentziano. Le misure di rumore sono piuttosto complesse e sono necessari setup di misura opportuni che consentano di amplificare la sorgente di rumore che si desidera analizzare così da poter ignorare tutti i contributi derivanti da altre sorgenti. Oltre all'analisi in termini di rumore, anche la caratterizzazione statica dei dispositivi elettronici riveste un ruolo fondamentale, non solo per la necessità di provvedere alla parametrizzazione del comportamento dei dispositivi quando i modelli forniti dalla fonderia appaiano carenti, ma anche in ambito industriale, ad esempio, per il collaudo, per il controllo di qualità e per la verifica delle proprietà del processo.

Questa tesi si occupa, nel dettaglio, della caratterizzazione statica e di rumore di dispositivi FinFET con lunghezza minima di canale di 14 nm.

Nel primo capitolo si discutono i principali fattori che ostacolano lo *scaling* delle tecnologie CMOS bulk, con un'analisi dettagliata dei limiti fisici e tecnologici. Successivamente si descrivono le due tecnologie FD-SOI CMOS e FinFET, che rappresentano un'evoluzione dei MOSFET bulk tradizionali.

Il secondo capitolo, dopo la descrizione dei dispositivi sotto misura (DUT, *Device Under Test*), tratta la caratterizzazione statica dei dispositivi a canale N e a canale P, presentando le curve tensione-corrente e discutendo l'estrazione di alcuni parametri, tra cui la transconduttanza g_m , l'efficienza di transconduttanza g_m/I_D , ed il guadagno di tensione intrinseco A_{Vi} .

L'ultimo capitolo riguarda la caratterizzazione in termini di rumore. Vengono richiamati, inizialmente, i modelli utilizzati per l'analisi di rumore dei dispositivi CMOS nanometrici e si discute nel seguito il setup di misura utilizzato. Il capitolo si conclude con la presentazione dei risultati di misura e l'estrazione e l'analisi dei parametri di rumore.

Capitolo 1

Tecnologie CMOS bulk e MOSFET a body ultra sottile

In questo capitolo, dopo una breve panoramica sul concetto di scaling delle tecnologie CMOS, si discuteranno i problemi indotti dalla continua miniaturizzazione delle dimensioni dei dispositivi, con particolare attenzione ai limiti fisici (*tunneling* nell'ossido ed effetti di canale corto), ai limiti tecnologici e dei materiali ed alle sfide che caratterizzano lo scaling dei tradizionali processi MOSFET e le corrispondenti innovazioni tecnologiche. Si prenderanno in considerazione alcune soluzioni candidate per la continuazione del processo di scaling dei dispositivi CMOS quali la tecnologia *Fully Depleted Silicon On Insulator, FD-SOI*, e la tecnologia FinFET, con enfasi sulle caratteristiche principali e sulle potenzialità in termini di controllo degli effetti di canale corto.

1.1 Scaling delle tecnologie CMOS

Negli ultimi decenni si è assistito ad un impressionante aumento della densità di integrazione e della complessità nei circuiti elettronici integrati. Il progredire del livello di integrazione, che fornisce una misura del numero di transistori realizzabili per unità di superficie di silicio, viene solitamente espresso tramite la legge di Moore, secondo la quale il numero dei componenti integrabili in una singola piastrina (o chip) subisce nel tempo una crescita esponenziale [1]:

$$N(y + n) = N(y) \cdot (1 + k)^n, \quad (1.1)$$

dove k indica l'incremento annuo del livello di integrazione (per esempio 0.5 per le memorie, 0.35 per i circuiti logici) ed $N(x)$ rappresenta il numero di com-

ponenti integrabili in una piastrina nell'anno x . La tecnologia microelettronica planare è alla base di tale evoluzione, in quanto ha permesso la miniaturizzazione dei dispositivi, che si traduce in una diminuzione dei consumi di potenza e dei costi, e in una costante riduzione delle minime dimensioni realizzabili, in particolare della lunghezza di canale minima di un transistor. A tal proposito, la figura 1.1 mostra l'evoluzione nel tempo della lunghezza minima di canale dei transistor MOS prevista dall'*International Technology Roadmap for Semiconductor* (nel 2014) [2].

A questo punto è lecito domandarsi come questa riduzione continua delle dimensioni influenzi le caratteristiche operative e le proprietà dei transistor MOS. Il processo di miniaturizzazione del dispositivo, infatti, è basato su una

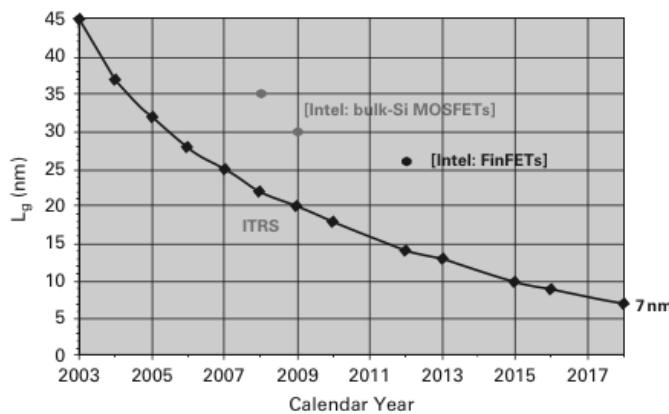


Figura 1.1: evoluzione della lunghezza minima di canale di transistori MOS nel tempo; i punti rappresentano i valori osservati o stimati.

metodologia che prevede l'impiego di alcune regole, denominate *regole di scaling*. Uno scaling improprio può produrre un aumento di alcuni effetti del secondo ordine che nascono o vengono amplificati nei MOSFET submicrometrici. Ad esempio, se si ipotizzasse di ridurre la sola lunghezza L del canale, mantenendo invariata la tensione applicata tra drain e source, il campo elettrico lungo il canale stesso diventerebbe più intenso, rendendo per esempio più marcato l'effetto dovuto ai portatori caldi. Risulta, dunque, di fondamentale importanza effettuare un corretto scaling del transistore in modo da stabilire come i parametri (geometrici e tecnologici) del MOSFET possano essere modificati per migliorare le prestazioni del dispositivo stesso. Un insieme completo di regole di scaling deve essere in grado di specificare come si modifica ciascun

parametro di un transistor quando si opera una riduzione delle dimensioni di un processo MOS preesistente. Tali regole non sono univoche, in quanto dipendono dalle proprietà che si vogliono conservare nel passaggio al MOSFET riscalato.

Il concetto di scaling è illustrato in tabella 1.1, dove si assume che le dimensioni del dispositivo nel passaggio da una generazione tecnologica alla successiva, scalino di un fattore α detto *dimensional scaling parameter*. Se, ad esempio, le dimensioni, il drogaggio e le tensioni sono scalate dello stesso fattore α , l'intensità dei campi elettrici nei dispositivi scalati è mantenuta costante, rispetto ai dispositivi originali. Questo scaling a campo costante, detto anche *scaling completo*, garantisce l'integrità dei dispositivi ed evita fenomeni di breakdown o altri effetti secondari.

Parametri fisici	Scaling a campo costante	Scaling generale	Scaling generale selettivo
Lunghezza di canale	$1/\alpha$	$1/\alpha$	$1/\alpha_d$
Spessore ossido	$1/\alpha$	$1/\alpha$	$1/\alpha_d$
Larghezza fili	$1/\alpha$	$1/\alpha$	$1/\alpha_w$
Larghezza canale	$1/\alpha$	$1/\alpha$	$1/\alpha_w$
Campo elettrico	1	ϵ	ϵ
Tensione	$1/\alpha$	ϵ/α	ϵ/α_d
Drogaggio	α	$\epsilon\alpha$	$\epsilon\alpha_d$
Area	$1/\alpha^2$	$1/\alpha^2$	$1/\alpha_w^2$
Capacità di gate	$1/\alpha$	$1/\alpha$	$1/\alpha_w$
Ritardo della porta	$1/\alpha$	$1/\alpha$	$1/\alpha_d$
Potenza dissipata	$1/\alpha^2$	ϵ^2/α^2	$\epsilon^2/\alpha_w\alpha_d$
Densità di potenza	1	ϵ^2	$\epsilon^2\alpha_w/\alpha_d$

Tabella 1.1: regole di scaling a campo costante, generale e selettivo.

L'operazione di scaling deve comunque far fronte ad alcuni problemi, ripresi, in dettaglio, nel paragrafo 1.1.1:

- il potenziale intrinseco (*built-in voltage*) delle giunzioni presenti nel dispositivo non scala con gli altri parametri in quanto legato all'intervallo proibito del silicio (*bandgap energy*), che non varia, a meno che non si utilizzi un semiconduttore differente (esclusi i casi di degenerazione del semiconduttore per elevati livelli di drogaggio e le variazioni dell'intervallo di energia proibita dovute a variazioni sostanziali della temperatura);

- la pendenza della corrente di drain come funzione della tensione tra gate e source in regione di sottosoglia dipende principalmente dalla distribuzione dell'energia dei portatori secondo Boltzmann e quindi non scala insieme agli altri parametri del dispositivo; pertanto la tensione di soglia non può essere ridotta eccessivamente senza produrre un aumento inaccettabile delle correnti di leakage e, di conseguenza, della dissipazione di potenza.

A causa di questi limiti può essere necessario un discostamento dalla semplice teoria di scaling a campo costante. In particolare quando la tensione di alimentazione è prossima ad 1 V, essa viene scalata di un fattore maggiore di $1/\alpha$, mediante l'introduzione di un parametro di scaling addizionale ϵ ($\epsilon > 1$) per il campo elettrico. Tale modello, definito di scaling generalizzato, viene indicato nella seconda colonna della tabella 1.1. Come si vede, l'aumento del campo elettrico rende necessario un incremento della densità di drogaggio (al fine di evitare il *punch-through* tra source e drain) e determina un aumento della densità di potenza dissipata. L'aumento del campo elettrico nel canale può inoltre portare problemi di affidabilità. Tuttavia si attenuano gli effetti sopracitati determinati dalla riduzione della tensione di alimentazione. Infine nella terza colonna della tabella 1.1 è riportato il cosiddetto *scaling selettivo*, che introduce un'ulteriore complicazione nel modello, prevedendo due parametri di scaling spaziale: α_d , applicato alle dimensioni verticali e alle lunghezze di gate, ed α_w , per lo scaling delle larghezze dei dispositivi e delle interconnessioni.

1.1.1 Teoria dello scaling a campo costante

Una delle teorie di scaling più note è quella dello *scaling a campo costante*, proposto da R. H. Dennard et al. in un articolo pubblicato nel 1974 [3].

Come già anticipato, questo metodo prevede di scalare le dimensioni del gate del MOSFET, intendendo per dimensioni sia la lunghezza di gate L_g sia la larghezza di gate W_g , e la tensione di alimentazione V_d dello stesso fattore α . Nella trattazione che segue si ipotizza che la lunghezza di canale L sia pari alla lunghezza L_g . Questo è vero approssimativamente nel caso in cui il transistor sia realizzato tramite il processo di autoallineamento. Si indicano, inoltre, le grandezze scalate con un apice (''): pertanto se L_g è la lunghezza di gate relativa al dispositivo originale, L'_g rappresenta la lunghezza di gate del dispositivo scalato. Sulla base di questa strategia e delle ipotesi fatte, il canale del transistor MOS scalato risulta caratterizzato dalle seguenti dimensioni geometriche:

$$L' = \frac{L}{\alpha}; \quad (1.2)$$

$$W' = \frac{W}{\alpha}. \quad (1.3)$$

Dello stesso fattore α sono ridotte anche le dimensioni verticali come lo spessore dell'ossido e le tensioni applicate al dispositivo, tra cui la tensione drain-source ($V'_{DS} = \frac{V_{DS}}{\alpha}$) e la tensione applicata tra gate e source ($V'_{GS} = \frac{V_{GS}}{\alpha}$). Infine la concentrazione di drogante è aumentata di un fattore α ($N'_A = \alpha \cdot N_A$). Questo principio viene illustrato nella figura 1.2, dove il MOSFET originale a canale N ha uno spessore d'ossido t_{ox} pari a 1000 \AA , un drogaggio di substrato N_A ed una polarizzazione scelti in modo che la tensione di soglia V_{TH} sia circa 2 V , e la lunghezza di canale L_g pari a $5 \mu\text{m}$. Applicando le regole di scaling viste con un fattore α pari a 5 si ottiene: $L'_g = 1 \mu\text{m}$, $t_{ox}' = 200 \text{ \AA}$ e $N'_A = 2.5 \cdot 10^{16} \text{ cm}^{-3}$. La strategia *a campo costante* si basa appunto sul principio di mantene-

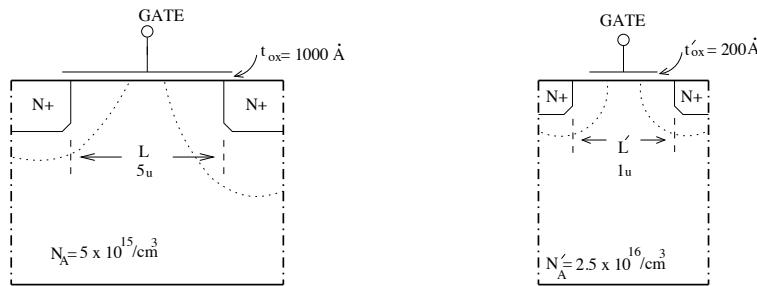


Figura 1.2: scaling di un NMOS con $\alpha = 5$.

re i campi nel dispositivo riscalato uguali a quelli presenti nel MOS originale. Mantenendo i campi costanti, infatti, molte delle caratteristiche del transistor con dimensione ridotta restano simili a quello del dispositivo prima dello scaling. A scopo di verifica, si analizza, dapprima, il campo orizzontale o longitudinale E_y .

Supponendo $V_{GS} > V_{TH}$ ed applicando una tensione V_{DS} , il campo elettrico longitudinale che si sviluppa sul canale di lunghezza L , risulta essere circa:

$$E_y \approx \frac{V_{DS}}{L}. \quad (1.4)$$

Pertanto, nel MOS scalato si ha:

$$E'_y \approx \frac{V'_{DS}}{L'} = \frac{V_{DS}}{\alpha} \cdot \frac{\alpha}{L} = E_y. \quad (1.5)$$

Dall'equazione (1.5) si nota che il fattore di riduzione applicato alla tensione consente di ottenere un campo longitudinale uguale a quello presente nel MOS

non scalato, fatto che dà il nome alla tecnica di scaling.

Ipotizzando ora di essere in regione di saturazione e di non eseguire per il momento alcuna variazione sulle grandezze tecnologiche che influenzano la tensione di soglia V_{TH} ed ammettendo, inoltre, che quest'ultima non dipenda dalla tensione tra source e body, la corrente di drain I_D del dispositivo scalato è data da :

$$I'_D = K' \cdot (V'_{GS} - V_{TH})^2. \quad (1.6)$$

Il fattore di corrente K' risulta uguale a K per l'ipotesi suddetta; infatti, se lo spessore dell'ossido t_{ox} rimane invariato e con lui anche la capacità dell'ossido C_{ox} si ha:

$$K' = \frac{\mu_n C_{ox} W'}{2L'} = \frac{\mu_n C_{ox} W}{2L} = K, \quad (1.7)$$

con μ_n mobilità degli elettroni. Conseguentemente, la corrente I_D diviene:

$$I'_D = K \cdot \left(\frac{V_{GS}}{\alpha} - V_{TH} \right)^2, \quad (1.8)$$

il che dimostra, che apportando solo una riduzione dei parametri V_{GS} , V_{DS} , L e W si avrebbe una riduzione notevole della corrente di drain, con conseguenti degradazioni delle prestazioni dinamiche del circuito. Analoghe considerazioni valgono per la corrente in triodo.

Al fine di mitigare il problema legato allo scaling della tensione tra gate e source si opera una riduzione della tensione di soglia V_{TH} intervenendo sui parametri tecnologici. La tensione di soglia dipende in modo particolare da due parametri di processo:

- concentrazione degli accettori nella regione di canale (N_A);
- spessore dell'ossido (t_{ox}).

La V_{TH} diminuisce al diminuire del valore di questi parametri. Una riduzione nel droggaggio del substrato N_A comporta tuttavia una crescita dei problemi relativi alla già menzionata *perforazione diretta*, a causa della quale le regioni di svuotamento del source e del drain si congiungono dando origine a un percorso di corrente sotto la regione di canale. Il problema, oltretutto, viene reso più acuto dalla riduzione della lunghezza di canale, che è uno dei risultati principali dell'operazione di scaling. D'altro canto, un aumento del droggaggio di substrato, pur risolvendo o, per lo meno, attenuando i problemi di *punch-through*, ne produrrebbe altri, quali l'aumento della capacità tra substrato e drain/source e la riduzione della tensione di rottura nelle giunzioni. Al fine di evitare questi effetti indesiderati, generalmente si esegue una impiantazione

superficiale di drogante nella regione di canale, impiantazione che viene tipicamente utilizzata per regolare la tensione di soglia dei dispositivi.

Risulta dunque evidente che, al fine di ridurre la tensione di soglia è necessario intervenire con una riduzione dello spessore dell'ossido di gate (t_{ox}). Questo, in aggiunta, provvede ad un aumento del fattore K , che rappresenta un secondo contributo all'incremento della corrente di drain. Si ha infatti che:

$$K' = \frac{\mu_n C'_{ox} W'}{2L'} = \frac{\mu_n C'_{ox} W}{2L} = \frac{\mu_n \epsilon_{ox} W}{2t'_{ox} L} = \frac{\alpha \mu_n \epsilon_{ox} W}{2t_{ox} L} = \alpha \cdot K, \quad (1.9)$$

dove si è espressa C_{ox} attraverso il rapporto ϵ_{ox}/t_{ox} (con ϵ_{ox} perennità dell'ossido). Per analizzare, infine, la variazione della tensione di soglia V'_{TH} dovuta all'aumento della concentrazione di drogante e alla riduzione dello spessore dell'ossido, si può esprimere V_{TH} come segue [4]:

$$V_{TH} = V_{FB} + 2|\phi_p| + V_C + \frac{1}{C_{ox}} \sqrt{2\epsilon_s q N_A (2|\phi_p| + V_C + V_B)}, \quad (1.10)$$

dove ϵ_s è la perennità dielettrica del silicio, ϕ_p è il potenziale nella regione p, V_C e V_B sono rispettivamente le tensioni di canale e di body, mentre V_{FB} rappresenta la tensione di banda piatta. Utilizzando, ora, le regole di scaling, riassunte nella prima colonna della tabella 1.1, nell'equazione 1.10 si ottiene:

$$V'_{TH} = \Phi'_{MS} - \frac{Q_f}{\alpha C_{ox}} + \frac{V_S}{\alpha} + 2|\phi'_p| + \frac{1}{\alpha C_{ox}} \sqrt{2\epsilon_s q N_A (2|\phi'_p| + \frac{V_{SB}}{\alpha})} \approx \frac{V_{TH}}{\alpha}, \quad (1.11)$$

dove si è espressa la tensione di banda piatta V_{FB} includendo la carica contenuta nell'ossido (con Φ_{MS} funzione lavoro metallo-semiconduttore e Q_f densità di carica fissa all'interfaccia) e si è riferita la tensione di soglia al source del dispositivo. Si è dunque trovato che anche la tensione di soglia viene ridotta approssimativamente secondo un fattore $1/\alpha$. Ciò considerato, la corrente di drain I_D in condizioni di saturazione diventa:

$$I'_D = K'(V'_{GS} - V'_{TH})^2 = \alpha K \cdot \left(\frac{V_{GS}}{\alpha} - \frac{V_{TH}}{\alpha} \right)^2 = \frac{K}{\alpha} (V_{GS} - V_{TH})^2 = \frac{I_D}{\alpha}. \quad (1.12)$$

Questa analisi, tuttavia, non tiene conto del fatto che variando il droggaggio del canale da N_A a αN_A si ha una riduzione della mobilità di canale, con la conseguenza che I_D diminuisce di un fattore superiore a $1/\alpha$.

La figura 1.3 illustra l'applicazione delle regole di scaling a campo costante per il MOSFET proposto in figura 1.2. Le figure sono basate sulle misure delle caratteristiche di un dispositivo reale e dimostrano una riuscita applicazione dello scaling proposto.

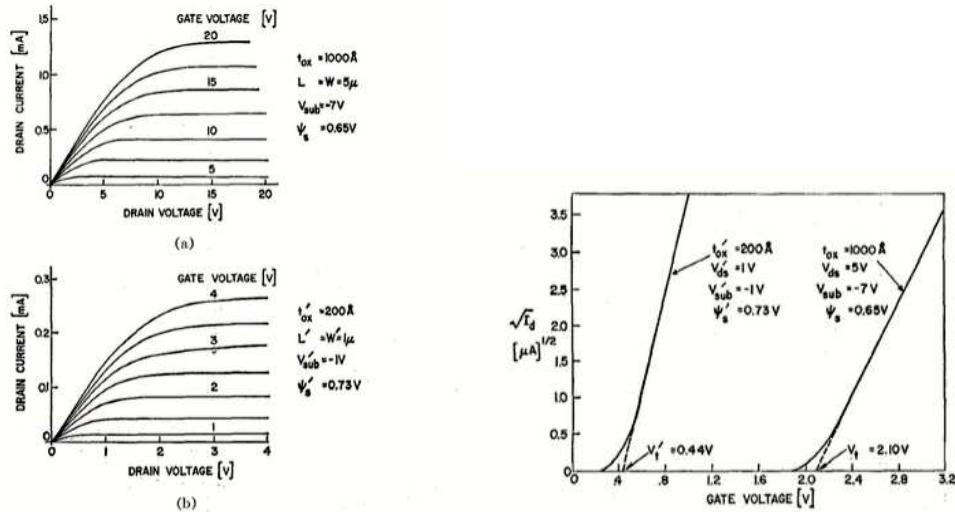


Figura 1.3: caratteristiche sperimentali $I_D - V_{DS}$ di un MOSFET originale (a) e di un MOSFET scalato (b) con $\frac{W}{L} = 1$ (a sinistra) e caratteristiche sperimentali di turn-on di un MOSFET originale e di uno riscalato con $\frac{W}{L} = 1$ (a destra) [3].

Effetti dello scaling a campo costante in logica CMOS

In questo paragrafo si analizzano gli effetti dello scaling della tecnologia sui parametri più importanti di un circuito digitale, come l'area, il ritardo e il consumo di potenza. Facendo riferimento al caso di un invertitore elementare, come quello mostrato in figura 1.4, che rappresenta un buon indicatore delle proprietà delle diverse porte logiche CMOS, lo scaling a campo costante influenza in modo positivo su gran parte dei parametri che lo caratterizzano. In modo particolare, l'area di gate dell'invertitore CMOS scalato è data da:

$$A'_{gate} = W'_n L'_n + W'_p L'_p = \frac{W_n L_n}{\alpha^2} + \frac{W_p L_p}{\alpha^2} = \frac{A_{gate}}{\alpha^2}, \quad (1.13)$$

e risulta pertanto ridotta del quadrato del fattore di scaling α .

Il tempo di propagazione di un inverter CMOS è espresso dalla seguente relazione [5]:

$$t_p = 0.52 \frac{C_L V_{DD}}{(W/L)_n K'_n V_{DSAT,n} (V_{DD} - V_{TH} - V_{DSAT,n}/2)}. \quad (1.14)$$

Utilizzando le regole di scaling a campo costante si trova facilmente che il

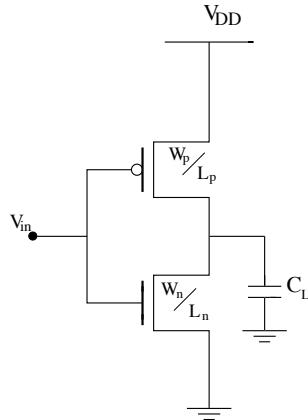


Figura 1.4: invertitore CMOS elementare.

tempo di propagazione scalato (t'_p) si riduce del fattore α . Infatti:

$$t'_p = 0.52 \frac{(C_L/\alpha)(V_{DD}/\alpha)}{(W/L)_n \alpha K'_n (V_{DSAT,n}/\alpha) (V_{DD} - V_{TH} - V_{DSAT,n}/2) (1/\alpha)} = \frac{t_p}{\alpha}. \quad (1.15)$$

Per quanto riguarda la potenza dissipata (dinamica), supponendo che essa sia dovuta alla sola carica della capacità parassita di uscita C_L , è possibile esprimerla mediante la relazione seguente:

$$P'_D = f' C'_L (V'_{DD})^2 = \alpha f \frac{C_L}{\alpha} \left(\frac{V_{DD}}{\alpha} \right)^2 = \frac{P_D}{\alpha^2}, \quad (1.16)$$

dove f rappresenta la frequenza delle commutazioni che può essere aumentata di un fattore α in quanto il tempo di propagazione t_p viene ridotto dello stesso fattore.

Limiti dello scaling a campo costante

Lo scaling a campo costante presenta problemi che possono essere riassunti nei punti seguenti:

- esso prevede la riduzione della tensione di alimentazione del medesimo fattore α utilizzato per ridurre delle dimensioni geometriche del dispositivo; tuttavia, in un circuito logico, ridurre la tensione di alimentazione può comportare una serie di problemi di compatibilità e con la componentistica preesistente;

- l'analisi condotta vale con buona approssimazione se tutte le tensioni in gioco scalano mediamente dello stesso fattore α ; dal momento che questo non è vero (ad esempio, le tensioni intrinseche di giunzione non scalano con gli altri parametri) nella realtà, i campi non rimangono costanti e le regioni di svuotamento sono molto ampie rispetto alle dimensioni fisiche del dispositivo, così da rendere più pronunciati gli effetti di canale corto;
- con il progredire dello scaling la tensione di soglia risulta essere estremamente bassa e conseguentemente diventa molto complicato spegnere il transistor; inoltre, la sua dipendenza dalla temperatura tende a diventare più significativa ($\approx -1 \frac{mV}{^\circ C}$ [6]).

1.1.2 Teoria generalizzata dello scaling

I risultati positivi dello scaling a campo costante mostrati nella figura 1.3 non implicano che il fattore α possa crescere arbitrariamente. Oltre ai già citati problemi che caratterizzano questo tipo di scaling esistono, infatti, dei limiti pratici, stabiliti dalle correnti di sottosoglia e dei vincoli posti dagli stessi materiali, come ad esempio la difficoltà di ottenere un ossido di gate sottile, uniforme e privo di imperfezioni. È stato dimostrato che per dimensioni al di sotto di 1 μm l'applicazione delle regole di scaling a campo costante producono risultati che si allontanano in maniera significativa dalle attese teoriche [6]. Sembra dunque opportuno modificare le regole di scaling in maniera tale che le tensioni di soglia e di alimentazione siano scalate in misura inferiore rispetto a quanto indicato nello scaling a campo costante. L'idea è quella di generalizzare la teoria di scaling a campo costante, consentendo al campo locale di aumentare, ma nello stesso tempo di conservare la forma e la distribuzione dei potenziali all'interno del dispositivo scalato [6].

Questo obiettivo viene raggiunto modificando le dimensioni fisiche del transistor e i potenziali applicati mediante fattori indipendenti, migliorando notevolmente in questo modo la flessibilità di progettazione. Per una data geometria del dispositivo ed un insieme di condizioni al contorno, l'equazione di Poisson e l'equazione di continuità per la corrente governano la configurazione del campo elettrico all'interno del MOSFET:

$$\frac{\partial^2 \phi}{\partial x^2} + \frac{\partial^2 \phi}{\partial y^2} + \frac{\partial^2 \phi}{\partial z^2} = -\frac{q}{\epsilon_s}(p - n + N_D + N_A); \quad (1.17)$$

$$\nabla J_n = 0, \quad (1.18)$$

dove $\phi(x)$ rappresenta il potenziale in x , p ed n la densità delle lacune e degli elettroni, q la carica elementare, N_D la concentrazione di donatori e J_n è la

denistà della corrente di elettroni. In condizioni di sottosoglia la concentrazione di elettroni contribuisce in modo trascurabile. Possiamo perciò trascurare il termine n nell'equazione (1.17). A questo punto le due equazioni (1.17) e (1.18) possono essere disaccoppiate (si noti infatti che J_n dipende solo dalla concentrazione di elettroni) e si può dunque trascurare la seconda.

Applicando le trasformazioni seguenti:

$$\phi' = \frac{\phi}{\alpha/\epsilon}; \quad (1.19)$$

$$(x', y', z') = \frac{(x, y, z)}{\alpha}; \quad (1.20)$$

$$(n', p', N'_A N'_D) = (n, p, N_D N_A) \epsilon \alpha, \quad (1.21)$$

nella (1.17), si ottiene:

$$\frac{\partial^2 \phi'}{\partial x'^2} + \frac{\partial^2 \phi'}{\partial y'^2} + \frac{\partial^2 \phi'}{\partial z'^2} = -\frac{q}{\epsilon_s}(p' - n' + N'_D + N'_A). \quad (1.22)$$

L'equazione (1.22) può essere interpretata come l'equazione di Poisson per un dispositivo scalato e, se i potenziali agli elettrodi di source, drain e gate sono ridotti proporzionalmente del fattore α/ϵ , le sue soluzioni differiscono da quelle della (1.17) solo per il fattore di scaling. In altre parole, la forma del campo elettrico è la stessa per i due dispositivi mentre la sua intensità varia come ϵ ; può dunque essere incrementata se $\epsilon > 1$. Le equazioni da (1.19) a (1.21) rappresentano le regole di *scaling generale*, riassunte nella seconda colonna della tabella 1.1.

Nonostante l'aumento dell'intensità del campo elettrico, la probabilità che si manifestino effetti di *punch-through* e di *Drain Induced Barrier Lowering*, trattati nel seguito, dovrebbe rimanere sostanzialmente invariata, in quanto viene conservata la distribuzione del campo elettrico.

La concentrazione di lacune e di elettroni è aumentata del fattore $\epsilon \alpha$, come mostrato nell'equazione (1.21). Questa assunzione può essere giustificata all'interno della regione di svuotamento, dove la densità di carica spaziale è influenzata solo dalla quantità di droggaggio e non dalla concentrazione di elettroni e lacune, che è funzione esponenziale del potenziale ϕ . All'interno di tale regione risulta infatti che $n, p \ll N_A$.

L'analisi fino ad ora condotta ha riguardato un MOS polarizzato in regione di sottosoglia. Se il transistor opera in forte inversione, non è più possibile trascurare l'equazione (1.18), in quanto la concentrazione di elettroni contribuisce alla carica spaziale ed a causa della non linearità dell'equazione del trasporto

non possiamo aspettarci che n scali del fattore $\alpha\epsilon$. Tuttavia, se lo strato di inversione è abbastanza sottile da trascurare i contributi legati al potenziale superficiale, la densità totale di elettroni per unità di area N_i scala ancora di $\alpha\epsilon$. In questo modo è possibile applicare le regole di scaling viste, anche nella condizione di forte inversione.

Con lo *scaling generale*, le prestazioni del dispositivo vengono migliorate dello stesso fattore rispetto al precedente scaling, mentre in termini di densità di potenza si ha un aumento di ϵ^2 . Si nota, infine, che se $\alpha = \epsilon$ si ricade nel caso dello *scaling a tensione costante*, in cui la tensione di alimentazione non viene variata, cosicché tutte le tensioni di polarizzazione relative al dispositivo scalato risultano essere identiche a quelle del dispositivo originale, mentre se $\epsilon = 1$ si hanno le regole dello scaling a campo costante. Questo modello rappresenta sicuramente un passo avanti rispetto alla teoria dello scaling a campo costante, in particolare dal punto di vista della compatibilità con le famiglie logiche CMOS appartenenti ai passati nodi tecnologici. D'altro canto, l'adozione del modello di scaling generalizzato comporta alcuni svantaggi come, ad esempio, l'amplificazione dei campi elettrici orizzontali e verticali nella regione di canale, che può dare luogo ad alcuni problemi e non idealità (per esempio, rottura dell'ossido di gate, degradazione della mobilità dei portatori dovuta al meccanismo di collisione superficiale e alla saturazione della velocità, etc.). Nel paragrafo successivo verranno discussi i principali limiti nello scaling delle tecnologie CMOS, in particolar modo dal punto di vista fisico e tecnologico.

1.2 Limiti nello scaling delle tecnologie CMOS bulk planari

La teoria base del transistor MOS contiene diverse approssimazioni il cui scopo è quello di semplificare la modellizzazione del comportamento del dispositivo. In realtà tale teoria è stata sviluppata in un periodo in cui i transistori MOSFET allo stato dell'arte avevano lunghezze di canale dell'ordine delle decine di micron (per questa ragione, la teoria è anche detta teoria del canale lungo). I limiti della teoria di base sono dunque diventati maggiormente evidenti con la continua miniaturizzazione dei dispositivi. Nel seguito si prenderanno in considerazione due tra le più importanti approssimazioni che caratterizzano la teoria base del transistore.

Definizione della tensione di soglia

Una delle più utili semplificazioni della teoria a canale lungo è quella di trascurare tutta la carica libera nel canale fino a quando la tensione di gate non

superi la tensione di soglia, che viene dunque definita considerando la carica nella regione svuotata come costante.

La carica totale indotta nel semiconduttore è la somma della carica mobile Q_n (carica nel canale) e della carica fissa Q_d (carica della regione di svuotamento):

$$Q_{TOT} = Q_n + Q_d. \quad (1.23)$$

Applicando la legge di Gauss¹ ad un volume che si estende dall'interfaccia ossido-silicio fino alla regione di volume priva di campo, la carica totale per unità di area risulta essere uguale a:

$$Q_{TOT} = -\epsilon_s E_{so}, \quad (1.24)$$

dove E_{so} rappresenta il campo nel semiconduttore all'interfaccia SiO_2-Si , che viene considerata priva di cariche libere e ϵ_s rappresenta la costante dielettrica del silicio. Mediante questa approssimazione, lo spostamento dielettrico D è continuo ed il campo E_{so} è legato a quello nell'ossido E_{ox} dal rapporto delle costanti dielettriche ϵ_{ox}/ϵ_s :

$$E_{ox}\epsilon_{ox} = E_{so}\epsilon_s. \quad (1.25)$$

Nell'ossido in assenza di cariche, il campo risulta costante e può essere espresso dalla seguente relazione [4]:

$$E_{ox} = \frac{V_{ox}}{x_{ox}} = \frac{1}{x_{ox}}[(V_G - V_B - V_{FB}) + (\phi_p - \phi_s)]. \quad (1.26)$$

Nell'equazione (1.26) il termine $(V_G - V_B - V_{FB})$ si riferisce alla caduta di potenziale effettiva che tende a caricare la struttura MOS mentre V_{ox} e $(\phi_p - \phi_s)$ rappresentano le tensioni rispettivamente ai capi dell'ossido e del silicio. Sostituendo la (1.26) nella (1.25), risolvendo rispetto a E_{so} e utilizzando la (1.24) otteniamo:

$$Q_{TOT} = -C_{ox}[(V_G - V_B - V_{FB}) + (\phi_p - \phi_s)], \quad (1.27)$$

da cui è possibile esprimere la carica mobile di canale come:

$$Q_n = -C_{ox}[(V_G - V_B - V_{FB}) + (\phi_p - \phi_s)] - Q_d. \quad (1.28)$$

Si può esprimere Q_n in funzione delle tensioni applicate notando che, in condizione di forte inversione, si ha:

$$\phi_s = -\phi_p + (V_C - V_B), \quad (1.29)$$

¹La legge di Gauss mette in relazione i campi su una superficie gaussiana (superficie chiusa di forma arbitraria) con le cariche racchiuse dalla superficie stessa.

e

$$Q_d = Q_{d,max} = -\sqrt{2\epsilon_s q N_A (2\phi_p + V_C - V_B)}, \quad (1.30)$$

da cui risulta che la carica mobile è data dalla relazione seguente:

$$Q_n = -C_{ox}[(V_G - V_C - V_{FB} - 2|\phi_p|) + \sqrt{2\epsilon_s q N_A (2\phi_p + V_C - V_B)}]. \quad (1.31)$$

Nella teoria del MOSFET a canale lungo, tutta la carica libera nel canale (Q_n) viene trascurata fino a quando l'ampiezza della tensione di gate non superi la tensione di soglia, cossicché la carica della regione svuotata (Q_d) viene considerata costante. In altre parole la tensione di soglia V_{TH} viene definita come la tensione di gate necessaria per indurre un canale conduttivo alla superficie del semiconduttore. Imponendo quindi $Q_n = 0$ nell'espressione (1.31), si ottiene la seguente equazione per V_{TH} :

$$V_{TH} = V_{FB} + V_C + 2|\phi_p| + \frac{1}{C_{ox}}\sqrt{2\epsilon_s q N_A (2\phi_p + V_C - V_B)}. \quad (1.32)$$

Nell'intorno della tensione di soglia, tuttavia, le correnti non sono descritte in maniera esauriente dalla teoria di base. Infatti, in questo caso, si osserva il passaggio di deboli correnti, definite correnti di sottosoglia, e l'approssimazione secondo la quale non vi sono elettroni alla superficie fino a quando non si raggiunge la condizione $V_{GS} > V_{TH}$ viene meno. La trattazione teorica della corrente di sottosoglia non viene qui illustrata in quanto nei MOSFET a canale nanometrico esistono diversi effetti che stanno diventando sempre più importanti e che vengono descritti nel seguito.

Approssimazione a canale graduale

Un'altra importante ipotesi utilizzata nella teoria del MOSFET a canale lungo, la cui validità si indebolisce al diminuire delle dimensioni del dispositivo, è l'approssimazione a *canale graduale*, utilizzata per dedurre l'equazione *a controllo di carica* [4] per la corrente di drain, che rappresenta, di fatto, l'espressione più utilizzata per analizzare il comportamento del MOS. In tale approssimazione si suppone che i campi nella direzione parallela al flusso di corrente (campi orizzontali) siano molto meno intensi di quelli nella direzione perpendicolare alla superficie del silicio (campi verticali) ($|\frac{\partial\phi}{\partial y}| \ll |\frac{\partial\phi}{\partial x}|$). Tale ipotesi giustifica l'uso di un'analisi unidimensionale del transistor per derivare le concentrazioni di portatori e le dimensioni della regione di svuotamento al di sotto del canale. In altre parole si ipotizza che la quantità di carica nel canale sia completamente controllata dall'elettrodo di gate (cioè dal campo perpendicolare all'interfaccia $Si-SiO_2$). Nei MOSFET di piccole dimensioni non è tuttavia possibile trascurare l'influenza delle giunzioni di drain e di source sulla quantità di carica nel

canale e la teoria di base sviluppata risulta inadeguata.

È infine importante considerare gli effetti legati ai forti campi a cui sono soggetti i MOS scalati, in quanto questi campi possono determinare effetti di portatori caldi, trascurabili, invece, nei dispositivi di grandi dimensioni.

L'inadeguatezza delle approssimazioni sopra discusse nel caso di dispositivi a canale corto, si manifesta in una serie di effetti che potevano essere ignorati nel caso di MOS con dimensione maggiore e diventano, invece, significativi nei dispositivi scalati, ponendo un limite allo scaling continuo del transistore CMOS convenzionale.

In generale questi effetti che limitano lo scaling dei dispositivi CMOS possono essere riassunti in 5 categorie:

- limiti fisici;
- limiti dei materiali;
- limiti tecnologici;
- limiti riguardanti la potenza termica dissipata;
- limiti economici.

1.2.1 Limiti fisici: tunneling nell'ossido di gate ed effetti di canale corto

Nella trattazione che segue verranno descritti i numerosi fenomeni di origine fisica che non possono essere ignorati nella realizzazione di dispositivi con dimensioni scalate.

Scaling dello spessore dell'ossido: oxide tunneling

Nel paragrafo 1.1.1 si è visto che lo scaling coinvolge anche lo spessore dell'ossido t_{ox} . Ad esempio, per un dispositivo CMOS con lunghezza di canale pari a 100 nm o meno è necessario uno spessore $< 3 \text{ nm}$ [7]; questo spessore comprende solo pochi strati di atomi e si avvicina ai limiti fondamentali collocati intorno a $1 - 1.5 \text{ nm}$.²

Uno strato così sottile è soggetto al meccanismo quantistico del *tunneling*, responsabile della generazione di una corrente di perdita che aumenta esponenzialmente con la riduzione dello spessore dell'ossido.

²Assumendo uno spessore $t_{ox} \approx 1.5 \text{ nm}$ e una tensione pari a 1 V ai capi dello strato di ossido, la densità di corrente di leakage di gate potrebbe raggiungere 10 A/cm^2 [7], [8]

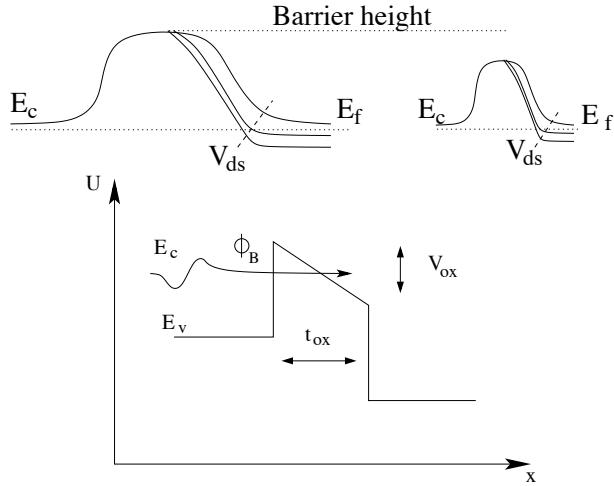


Figura 1.5: riduzione della barriera di energia che ostacola il flusso da source a canale dei portatori maggioritari dovuta allo scaling a campo costante (*in alto*), approssimazione di una barriera di potenziale triangolare utilizzata nella trattazione, dove E_V è l'energia del livello di valenza, E_C quella del livello di conduzione e t_{ox} rappresenta lo spessore della barriera (*in basso*).

Questa relazione tra la corrente di perdita e lo spessore dell'ossido può essere trattata considerando il diagramma a bande del dispositivo scalato in figura 1.5.

Ipotizzando una barriera di potenziale triangolare, la probabilità di attraversamento J da parte di un elettrone si può esprimere come:

$$J = J_0 E_{ox}^2 e^{-\frac{B}{E_{ox}}}, \quad (1.33)$$

dove $E_{ox} = \frac{V_{ox}}{t_{ox}}$ rappresenta il campo elettrico applicato alla barriera (campo sull'ossido), V_{ox} la differenza di tensione che si distribuisce sull'ossido e $B = f(\phi_B)$ è funzione del potenziale elettrostatico tra l'apice della barriera e il livello di conduzione. Con il diminuire di t_{ox} tale probabilità aumenta esponenzialmente, pregiudicando ovviamente le proprietà di isolamento del dispositivo.

In figura 1.6 sono rappresentate le correnti dovute al tunneling per spessori di ossido che variano da 3,6 nm a 1 nm in funzione della tensione di gate. La figura mostra che per uno spessore di ossido pari a 20 Å, le correnti di perdita possono salire fino a circa $1 - 10 \text{ A/cm}^2$ [9]. Queste correnti relativamente elevate possono alterare le prestazioni del dispositivo, in particolare dal punto di vista della potenza dissipata.

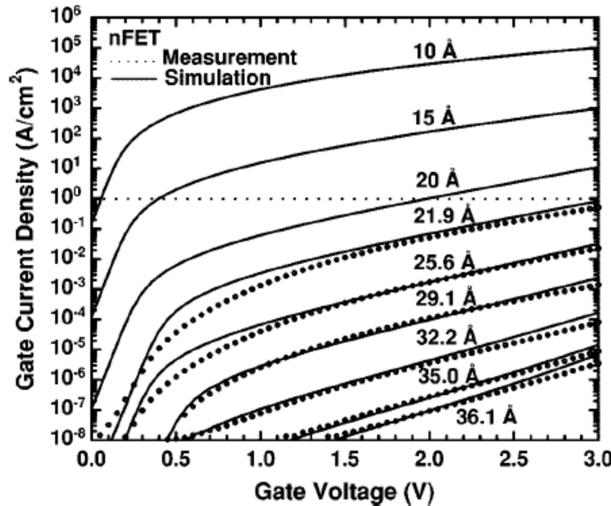


Figura 1.6: correnti di perdita dovute al *tunneling* in funzione della tensione di gate per differenti spessori di ossido [9].

Si può pensare di limitare i leakage di corrente utilizzando ossidi più spessi; questo però comporterebbe un degrado della capacità dell'ossido C_{ox} , definita come:

$$C_{ox} = \frac{\epsilon_{ox}}{t_{ox}} = \frac{\kappa \cdot \epsilon_0}{t_{ox}}, \quad (1.34)$$

e, conseguentemente, della *driving capability* del transistore. Appare evidente dalla (1.34) che se si aumenta lo spessore dell'ossido di gate al fine di limitare la corrente di leakage, il valore di C_{ox} può essere mantenuto invariato aumentando la costante dielettrica κ . L'idea è, quindi, quella di sostituire il biossido di silicio con nuovi materiali ad elevato κ , come, ad esempio Ta_2O_5 , TiO_2 , ZrO_2 . La ricerca è, comunque, focalizzata su materiali con ampio *energy gap* ed alta barriera di potenziale, in quanto le correnti di *leakage* dipendono esponenzialmente anche da tali fattori³ [10].

Scaling della lunghezza di canale: effetti di canale corto

I problemi indotti dalla continua miniaturizzazione delle dimensioni dei dispositivi, ed in particolar modo, della lunghezza di canale L , vengono denominati *effetti di canale corto*. In generale tali effetti determinano un peggioramento della transconduttanza g_m del dispositivo, in quanto il controllo della tensione

³L'introduzione dei materiali ad alta costante dielettrica è vista come unica prospettiva dalla ITRS per spessori equivalenti di ossido che raggiungono 0.5 nm.

di gate V_G sulla corrente di drain tende ad indebolirsi.

Come detto all'inizio del paragrafo, nella teoria base del transistore, per la formulazione dell'equazione a controllo di carica si ricorre all'approssimazione a canale graduale, che nel caso di MOS con lunghezze di canale corto non è più soddisfatta, in quanto le giunzioni *pn* presenti al source e al drain influenzano la regione di carica spaziale nel canale. Una teoria monodimensionale risulta, difatti, inadeguata per tenere conto dello scaling tecnologico, rendendo quindi necessaria un'analisi bidimensionale che coinvolga sia i campi longitudinali sia quelli trasversali. In termini del tutto generali il valore della tensione di soglia trovato sulla base della teoria del transistor a canale lungo risulta essere elevato rispetto a quello reale di un MOSFET a canale corto in quanto si ipotizza che tutta la carica nella regione svuotata sia imputabile al solo gate. Questo significa che nel caso di MOS a canale corto la tensione di gate V_G necessaria per indurre una data carica nel canale risulta più bassa rispetto al caso di un MOS a canale lungo, il che equivale a dire che la tensione di soglia V_{TH} si riduce al diminuire della lunghezza di canale.

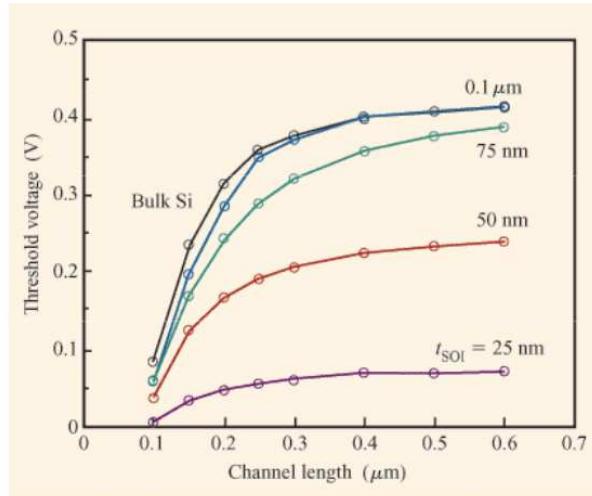


Figura 1.7: variazione della tensione di soglia V_{TH} in funzione della lunghezza L di canale [11].

La figura 1.7 mostra la dipendenza della tensione di soglia rispetto alla lunghezza L di canale per diversi nodi tecnologici (incluso uno relativo ad un processo *Silicon On Insulator, SOI*) e permette di evidenziare come V_{TH} tenda a decrescere al ridursi di L e come tale diminuzione diventi estremamente rapida per valori di L inferiori a $0.2 \mu\text{m}$.

La diminuzione della tensione V_{TH} può essere vista, inoltre, come conseguenza della riduzione dell'altezza della barriera di potenziale, mostrata in figura 1.5. All'aumentare della tensione di polarizzazione al terminale di drain, le due giunzioni pn *body – source* e *body – drain* possono arrivare a sovrapporsi; conseguentemente, i portatori maggioritari iniziano a fluire dal source al drain anche se la tensione di gate V_G è al di sotto del valore di soglia, determinando gli effetti di *perforazione diretta* ed il fenomeno del *Drain Induced Barrier Lowering, DIBL*, trattato nel seguito.

Drain Induced Barrier Lowering. In debole inversione si ha una barriera di potenziale tra il source e la regione di canale, la cui altezza è il risultato del bilanciamento tra la corrente di drift e quella di diffusione tra le due regioni. La barriera, idealmente, dovrebbe essere controllata dall'elettrodo di gate. Tuttavia essa viene influenzata dal campo elettrico presente al drain quando si applica una differenza di potenziale V_{DS} tra drain e source. In modo particolare, per *Drain Induced Barrier Lowering* si intende l'abbassamento della barriera di potenziale, ad opera di una tensione applicata all'elettrodo di drain. Tale effetto rende la tensione di soglia dipendente dalle tensioni operative, come mostrato in figura 1.8, per tensioni basse di drain.

All'aumentare della tensione V_{DS} , la regione di svuotamento di drain, infatti, si muove verso la regione di svuotamento del source. Conseguentemente si assiste ad una penetrazione del campo elettrico di drain nella regione di canale, normalmente controllata dal gate. A causa di ciò, la barriera di potenziale al source si abbassa, portando ad un incremento di elettroni iniettati dal source, al di sopra di tale barriera nel canale, e quindi ad una diminuzione della tensione di soglia.

L'influenza dell'effetto di DIBL nei dispositivi scalati può essere analizzato risolvendo l'equazione di Poisson bidimensionale numericamente. Tuttavia la

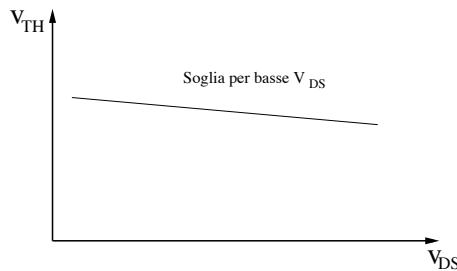


Figura 1.8: variazione della tensione di soglia V_{TH} per bassa tensione di polarizzazione V_{DS} (Drain Induced Barrier Lowering) [4].

tensione di soglia che ne risulta è troppo complessa per essere utilizzata nei simulatori circuitali. Esiste, però, un semplice modello [12] (mostrato più avanti) che risulta estremamente accurato per i dispositivi MOS scalati con bassa tensione di polarizzazione al drain ($V_{DS} < 0.1$ V).

Per valori sufficientemente elevati della tensione di drain V_{DS} , le regioni di source e drain possono trovarsi addirittura cortocircuitate, provocando un brusco aumento della corrente. Tale effetto, definito *punch-through*, stabilisce un limite superiore alla tensione di drain che può essere applicata al dispositivo.

Effetti di portatori caldi. Come detto precedentemente, nello scaling generalizzato le dimensioni dei dispositivi e le tensioni di alimentazione ed operative vengono ridotte mediante fattori di scaling differenti. Il conseguente aumento dell'intensità del campo elettrico comporta l'accelerazione degli elettroni che possono raggiungere livelli di energia cinetica ben al di sopra di quella termica tipica dei portatori all'equilibrio termodinamico. Questi elettroni prendono il nome di portatori caldi. I meccanismi principali con cui essi interagiscono con la struttura MOS sono:

- iniezione di portatori nello strato di ossido (*per effetto tunnel*);
- generazione di coppie elettrone-lacuna per effetto valanga;
- iniezione di portatori dal source verso il substrato.

Se gli effetti analizzati precedentemente comportano una riduzione della tensione di soglia all'interno dello stesso sistema, questi determinano derive nel tempo e possono provocare un aumento o una riduzione della V_{TH} (si pensi ad esempio agli elettroni intrappolati nello strato di ossido che provocano una aumento della tensione di soglia nei dispositivi NMOS e una diminuzione nei transistori PMOS).

Random dopant fluctuations (RDF). La tensione di soglia V_{TH} può subire delle variazioni statistiche che possono essere sostanzialmente di due tipi: statiche o dinamiche. Le prime, in particolar modo, sono caratterizzate da diversi fattori, come variazioni del processo di fabbricazione, fluttuazioni nelle geometrie, intrappolamento non controllato di carica nell'ossido, ma soprattutto variazioni di drogante sia in termini di concentrazione sia di posizione (*Random dopant fluctuations*). Le fluttuazioni casuali di drogante sono introdotte dal processo di impiantazione e dalla conseguente diffusione degli atomi droganti. Durante l'impiantazione ionica, gli atomi droganti sono impiantati all'interno del reticolo cristallino con un'adeguata energia e vengono attivati

mediante *annealing*, che consente alle impurità di occupare le corrette posizioni sostituzionali nel reticolo cristallino. Tuttavia gli atomi droganti possono collocarsi in posizioni casuali all'interno della regione attiva del dispositivo in quanto, durante il processo mediante il quale si introducono le impurità, si possono verificare una serie di collisioni casuali. Queste alterano le caratteristiche elettriche da un dispositivo all'altro, introducendo fluttuazioni della tensione di soglia, che diventa sempre più pronunciata nei dispositivi caratterizzati da una piccola lunghezza di gate a causa della riduzione nel numero medio di atomi drogati, come illustrato nella figura 1.9. Conseguentemente, dispositivi nominalmente identici risultano avere caratteristiche traslate tra loro e perciò soglie differenti.

Per ridurre l'impatto dell'RDF sulle caratteristiche elettriche si adottano particolari profili di droggaggio che, comunque, non eliminano completamente il problema. Come si vedrà nel seguito, nei dispositivi UTB SOI l'effetto di RDF è ridotto mediante la realizzazione di un canale debolmente drogato.

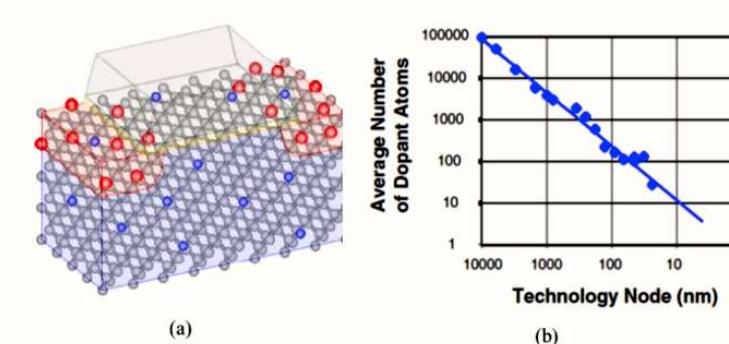


Figura 1.9: a) schema di un MOSFET sotto l'influenza di RDF. I punti rossi e blu rappresentano gli atomi donatori e accettori, mentre i punti grigi sono gli atomi di silicio del reticolo cristallino [13]; b) numero medio di atomi droganti in funzione del nodo tecnologico [14].

Analisi della riduzione della tensione di soglia

Abbiamo visto come l'espressione della tensione di soglia V_{TH} formulata per i MOS a canale lungo risulti inadeguata per l'analisi dei dispositivi scalati. Esiste una tecnica geometrica [12] che permette di derivare, in modo approssimato, l'espressione di V_{TH} per dispositivi con lunghezza di canale dell'ordine di 1 μm che si accorda abbastanza bene con i valori sperimentali.

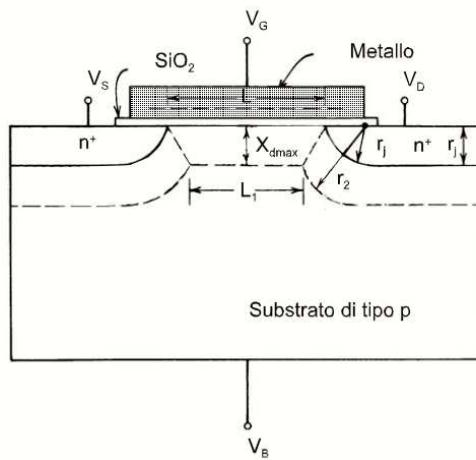


Figura 1.10: sezione trasversale di un MOS a canale corto che illustra il modello geometrico.

La figura 1.10 illustra il modello geometrico che spiega la riduzione della tensione di soglia, prevista dagli effetti di canale corto. Per piccoli valori di V_{DS} , si considera la carica indotta da V_G contenuta approssimativamente in un volume la cui sezione trasversale è un trapezio di altezza x_{dmax} , base maggiore L e base minore L_1 , come mostrato in figura. Scegliendo la larghezza di canale pari a W e la concentrazione di drogante uguale a N_A , a tale regione di carica spaziale è associata una Q_{d1} pari a:

$$Q_{d1} = qx_{dmax}WN_a \frac{L + L_1}{2}, \quad (1.35)$$

che rappresenta la carica che deve essere indotta dal gate nella regione di carica spaziale per portare il canale alla soglia di conduzione (nella teoria di base $Q_{d1} = Q_d = qx_{dmax}N_aWL$, in quanto $L = L_1$). Nei MOS a canale corto, $L_1 < L$ e quindi Q_{d1} risulta minore della carica indotta dal gate in condizioni di canale lungo, come previsto in base alle precedenti considerazioni qualitative. Esprimendo L_1 in termini relativi alla geometria del MOS, in particolare utilizzando l'approssimazione secondo cui sempre in figura 1.10, $r_2 = r_j + x_{dmax}$, e mediante considerazioni di natura geometrica, si conclude che:

$$f = \frac{Q_{d1}}{Q_d} = 1 - \frac{r_j}{L} \cdot \left(\sqrt{1 + \frac{2x_{dmax}}{r_j}} - 1 \right). \quad (1.36)$$

L'equazione (1.36) rivela che il rapporto tra la carica indotta dal gate nel caso di canale corto e quella nel caso di canale lungo è unicamente determinato dalla geometria del MOS. In accordo con la (1.10), l'equazione della tensione di soglia può essere espressa da:

$$V_{TH} = V_{FB} + 2|\phi_p| + V_S + \frac{f}{C_{ox}}\sqrt{2\epsilon_s q N_a (2|\phi_p| + V_S + V_B)}, \quad (1.37)$$

dove, in questo caso, la V_{TH} è riferita al source. Questa formula è ampiamente utilizzata ed approssima in maniera corretta i dati sperimentali trovati.

1.2.2 Limiti tecnologici e dei materiali

Lo scaling dei dispositivi non comporta solamente sfide in ambito fisico, con l'introduzione quindi di nuove soluzioni e passi di processo addizionali per ridurre l'entità dei problemi esaminati, ma mette anche limiti associati ai materiali utilizzati nel processo produttivo e limiti tecnologici. Quest'ultimi, in modo particolare, sono dovuti alla barriera imposta dalle tecniche litografiche, che non consentono di ottenere risoluzioni molto migliori della lunghezza d'onda utilizzata nel processo litografico. Quando le dimensioni minime del dispositivo da integrare diventano comparabili con la lunghezza d'onda della luce impiegata nei sistemi ottici di esposizione, infatti, i fenomeni di diffrazione possono limitare la risoluzione ottenibile. Esistono, comunque, delle tecniche di esposizione innovative [15], ancora oggetto di studio, che consentono di superare questo limite. Una soluzione semplice è quella di utilizzare come sorgente di illuminazione per l'esposizione del *resist*, una luce ultravioletta di lunghezza d'onda più corta. I vantaggi di questa tecnica sono tuttavia modesti e, pertanto, per spingere al limite le possibilità dei processi litografici, si possono adottare delle tecniche che impiegano fasci di elettroni, raggi X o fasci di ioni. In particolar modo, nella litografia a fascio elettronico (*electron-beam lithography*) la fetta è ricoperta di un resist sensibile ai fasci elettronici (elettroresist) e viene esposta, senza l'interposizione di una maschera, all'azione di un fascio elettronico ben collimato controllato da elaboratore che viene deflesso per riprodurre la geometria desiderata. Il problema maggiore di questa tecnica litografica è la bassa produttività a causa della sequenzialità dell'operazione e della scarsa sensibilità dell'elettroresist. Uno sviluppo più recente è costituito dall'impiego della litografia a raggi X, in cui il fascio viene fatto passare attraverso una maschera per impressionare uno strato di resina fotosensibile. La litografia a raggi X, come la fotolitografia, impressiona simultaneamente le diverse geometrie e quindi permette la produzione su larga scala. I principali limiti sono dovuti alle specifiche richieste per le sorgenti di raggi X e per la

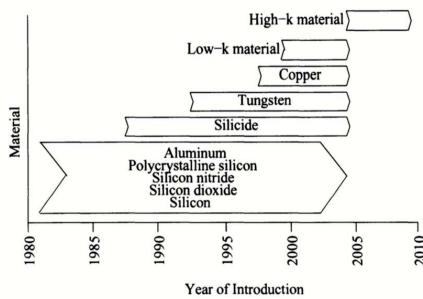


Figura 1.11: progressiva introduzione di nuovi materiali nel processo produttivo in tre decadi, dal 1980 al 2010 [10].

fabbricazione delle maschere. Un serio inconveniente nell'utilizzo dei raggi X è che l'estensione dell'area esposta in un intervallo di tempo non può superare certi limiti, per via della deformazione che la fetta subisce durante il processo. Inoltre è necessario considerare i danni che i raggi ad alta energia possono provocare alle regioni attive dei dispositivi nel silicio.

In figura 1.11 sono mostrati i nuovi materiali introdotti, lungo tre decadi, nei processi produttivi microelettronici, al fine di tenere il passo con il continuo ridimensionamento dei transistor CMOS. I materiali utilizzati nel processo di produzione CMOS sono diversi ed includono ad esempio, silicio, ossido di silicio, rame, alluminio. Con il continuo scaling del dispositivo questi materiali non possono più garantire, ad esempio, un isolamento affidabile o una buona conduzione. Di fronte a queste limitazioni nasce, dunque, la necessità di ricercare nuovi materiali per sostituire quelli convenzionali, per esempio materiali ad elevata costante dielettrica in grado di rimpiazzare il biossido di silicio come strato isolante tra gate e canale. Tuttavia questi non sono in realtà immuni da problemi; per esempio, alcuni di essi hanno la tendenza a cambiare le loro proprietà con la variazione della temperatura.

Al fine di sfruttare al massimo le tecnologie CMOS tradizionali esistono altre tecniche che consentono di migliorare le prestazioni dei dispositivi, ad esempio in termini di mobilità dei portatori. A tal proposito è importante citare le tecniche di *strained silicon* [16], ed in particolare la tecnica, proposta da Intel, che si basa sull'adozione del silicio-germanio. Questa tecnica consiste nel depositare sul corpo del wafer di silicio uno strato di silicio-germanio di spessore pari a circa $2 \mu\text{m}$ e concentrazione di germanio pari al 20%; la concentrazione del germanio non è uniforme su tutto lo strato ma risulta maggiore quanto più ci si allontana dall'interfaccia con il wafer originario. A questo punto un sotti-

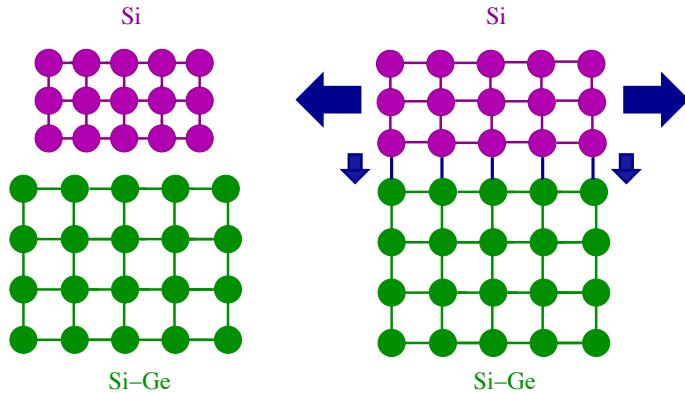


Figura 1.12: reticolo di atomi di "silicio stirato".

lissimo strato di silicio di spessore di circa 20 nm viene depositato sullo strato di Si-Ge. In questo modo gli atomi di silicio dello strato sovrastante tendono ad allinearsi con quelli dello strato Si-Ge che, essendo più spesso, obbliga gli atomi di silicio a separarsi di una distanza analoga a quella degli atomi di silicio-germanio, come mostrato in figura 1.12. In questo modo il reticolo cristallino del silicio viene allungato di circa l'1%, sia in direzione laterale che verticale, permettendo un incremento della mobilità dei portatori di carica che incontrano una resistenza inferiore al loro passaggio.

I vari problemi illustrati sino ad ora sono andati aggravandosi con l'evoluzione delle tecnologie CMOS ed appare oggi ormai assodato che i processi CMOS bulk convenzionali non saranno in grado di soddisfare in maniera adeguata la legge di Moore. Per questa ragione, nuove tecnologie, che introducono sostanziali variazioni rispetto a quelle tradizionali e sembrano in grado di mantenere il passo con la legge di Moore, sono da tempo allo studio, quando non si siano già affacciate sul mercato. Nei paragrafi che seguono verranno esaminate le caratteristiche di due nuove tecnologie considerate tra le più promettenti, entrambe basate sull'impiego di un body ultra sottile (*Ultra Thin Body, UTB*).

1.3 Tecnologie MOSFET a body ultra sottile

La riduzione continua delle dimensioni dei MOSFET e di conseguenza l'accentuarsi dei limiti che essa comporta impone il ricorso a nuove tecnologie, che prevedano l'introduzione sia di nuovi materiali, sia di strutture MOS avanzate

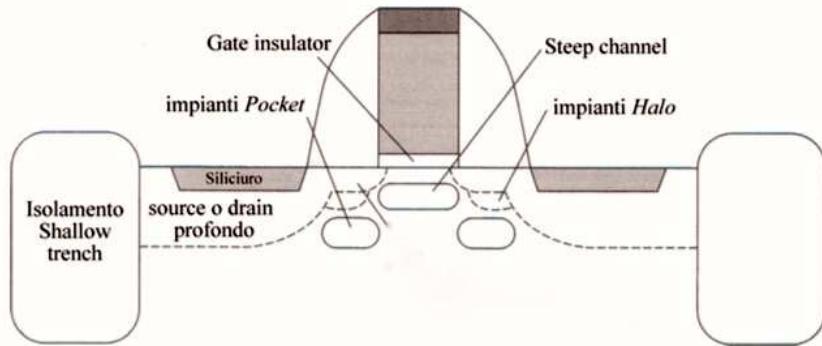


Figura 1.13: sezione trasversale di un MOSFET nanometrico.

e non convenzionali. In questo paragrafo ci si concentra sui cosiddetti *"Non-classical MOSFETs"* [11], che rappresentano l'evoluzione dei MOSFET bulk tradizionali, e cercano di fronteggiare e risolvere i diversi problemi analizzati precedentemente.

Le principali tecnologie, candidate per la continuazione del processo di scaling dei dispositivi CMOS, sono:

- la tecnologia *Fully Depleted SOI, (FD-SOI)* trattata nel paragrafo 1.3.1,
- la tecnologia *FinFET*, studiata nel paragrafo 1.3.3.

In un MOS convenzionale i limiti di canale corto vengono, generalmente, controllati mediante un preciso profilo di drogaggio che coinvolge soprattutto la regione superficiale di canale. Sagomare in maniera opportuna il drogaggio sia in profondità che in lunghezza del canale è possibile ma nello stesso tempo è molto complesso e costoso.

In figura 1.13 viene mostrata, schematicamente, la sezione trasversale di un MOS nanometrico, in cui si evidenziano le diverse soluzioni tecnologiche adottate. In particolare gli *strati di silicio* sulle regioni di source e drain permettono di ridurre le resistenze parassite (soluzione adottata nelle tecnologie ad altissima densità di integrazione). Le diffusioni di source e drain vengono ottenute mediante tecniche di *lightly doped drain*, che consistono nella realizzazione di una regione a basso drogaggio al confine col canale al fine di ridurre il valore del campo elettrico all'interfaccia tra canale e drain. La presenza dello *steep channel* crea un gradiente di concentrazione di cariche positive sogrammando il campo elettrico nel canale in modo tale da allontanare gli elettroni

dalla superficie e, in tal modo, diminuire i fenomeni di scattering e aumentare la mobilità dei portatori. Gli *halo implant* schermano il canale dalle regioni di source e di drain più profonde, limitando quindi gli effetti di canale corto. Infine, i *pocket implant*, riducono il fenomeno di DIBL, in quanto diminuiscono l'estensione dello svuotamento di source e drain.

Queste sono alcune delle soluzioni tecnologiche che vengono adottate nella realizzazione di MOSFET submicrometrici. Come si può notare il processo è sempre più complicato e dispendioso.

Come vedremo nel dettaglio, le nuove tecnologie, sopra menzionate, si basano su un body ultra sottile (*Ultra-Thin Body*, UTB) per controllare gli effetti legati alla riduzione della lunghezza di canale; inoltre il *FD-SOI MOSFET* è caratterizzato da un solo gate (anche se il substrato può essere considerato come un secondo gate), mentre il *FinFET* può prevedere l'impiego di due o tre gate (si parla di *DG-FinFET* o *TG-FinFET*, dove *DG* e *TG* stanno, rispettivamente, per *double gate* e *triple gate*).

1.3.1 Tecnologie fully depleted SOI

La struttura base di un SOI MOSFET è quella rappresentata in figura 1.14. Il processo di produzione è molto simile a quello di un convenzionale bulk-CMOS, tranne che per la presenza di un UTB deposto sul *Buried Oxide* (BOX), che isola il sottile strato di Silicio dal substrato. In un *Fully Depleted (FD) SOI MOSFET*, lo spessore del film di silicio ($\approx 10 \text{ nm}$) è inferiore alla profondità massima dello strato di svuotamento. Pertanto l'intera regione di substrato risulta svuotata di portatori maggioritari sotto qualsiasi condizione di polarizzazione; se invece lo spessore di Silicio al di sopra del BOX è abbastanza spesso da permettere la presenza di un body quasi neutro (non del tutto svuotato), si parla di *Partially Depeted (PD) SOI MOSFET*.

La presenza di un substrato completamente svuotato comporta proprietà molto interessanti e utili, come un'alta transconduttanza, un basso campo elettrico ed una pendenza di sottosoglia molto ripida. Tuttavia, in dispositivi submicrometrici, le prestazioni peggiorano a causa degli effetti di canale corto.

L'aumento, in particolare, dell'effetto di DIBL può essere mitigato con la realizzazione di un BOX sottile (in quanto, come vedremo più avanti, ridurre lo spessore del BOX facilita la soppressione dell'effetto di *fringing field* del drain) e mediante la realizzazione di un piano di massa (*ground plane (GP)*), come mostrato in figura 1.15.

Il concetto di piano di massa [17], [18] è, infatti, una delle tecniche utilizzate per ridurre gli effetti di canale corto ed è maggiormente efficace quando la distanza tra il GP e il drain risulta piccola rispetto alla lunghezza di canale. Se,

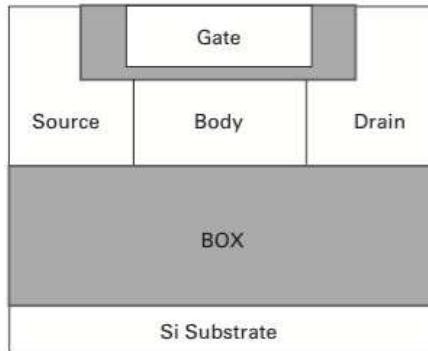


Figura 1.14: sezione trasversale di base di un SOI MOSFET [11].

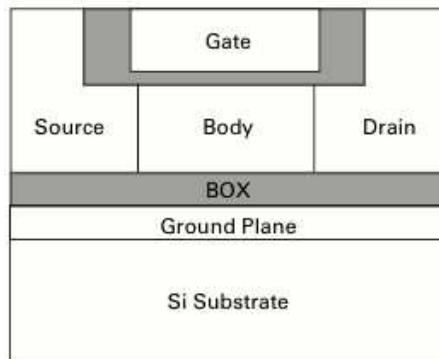


Figura 1.15: sezione trasversale di base di un SOI MOSFET con BOX sottile e GP [11].

come mostrato nella figura 1.15, il piano di massa è posizionato nel substrato (*Ground Plane in Substrate (GPS)*) risulta necessario avere un BOX il più sottile possibile, il che comporta un'aumento della pendenza della caratteristica corrente-tensione nella regione di sottosoglia. Esistono tuttavia dei processi in cui il piano di massa viene introdotto nel BOX (*Ground Plane in Buried Oxide (GPB)*) per migliorare gli effetti legati alla riduzione del canale, sia dal punto di vista del DIBL sia da quello della tensione di soglia. La figura 1.16 mostra le tre differenti strutture: a) FD SOI MOSFET convenzionale, b) FD SOI MOSFET con GPS, c) FD SOI MOSFET con GPB.

Il ridotto spessore del body nei MOSFET FD-SOI consente, inoltre, un accoppiamento elettrico tra il gate ed il substrato (che, come già accennato, può essere visto come secondo gate). Per questo, generalmente, si utilizza anche

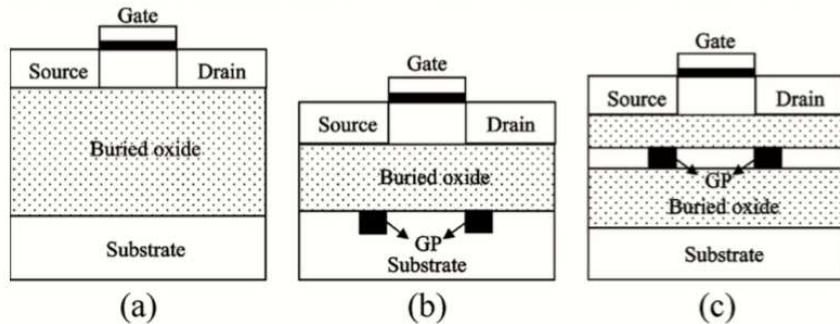


Figura 1.16: sezione trasversale delle tre strutture FD SOI: (a) FD SOI MOSFET convenzionale, (b) FD SOI MOSFET con GPS, (c) FD SOI MOSFET con GPB [17].

un BOX sottile che migliora tale accoppiamento rendendo V_{TH} dipendente in maniera significativa dal drogaggio del substrato, dalla sua polarizzazione nonchè dagli spessori dell'UTB (t_{SI}) e del BOX (t_{BOX}). Nell'analisi che segue si discutono le principali caratteristiche del FD-SOI MOSFET caratterizzato da un ground plane posizionato nel substrato, come illustrato schematicamente nella figura 1.15.

Uno degli aspetti da chiarire è relativo all'utilizzo di un BOX sottile. In un SOI MOSFET dotato di un BOX spesso, come mostrato in figura 1.14, il campo elettrico trasversale nell'isolante viene sopraffatto dal campo elettrico (*fringing field*) del source e del drain. Infatti, quando la lunghezza del canale è comparabile con lo spessore della regione di svuotamento della giunzione pn formata dalle regioni di drain/source e quella di substrato, il campo di *fringing* tende ad aumentare gli effetti di canale corto e la corrente di sottosoglia. La figura 1.17 illustra la distribuzione di potenziale elettrostatico per un dispositivo caratterizzato da un BOX spesso e per uno caratterizzato da un BOX sottile, in condizioni di bassa corrente di polarizzazione di drain. Come si nota, le linee del campo elettrico nel BOX partono per lo più dalle regioni di drain e di source e tendono a terminare nella regione di canale. Conseguentemente si ha un forte effetto di *fringing field*. Nel dispositivo realizzato con BOX sottile, invece, le linee del campo tendono a terminare nel substrato portando ad una riduzione degli effetti di fringing. Sulla base delle analisi condotte per la quantificazione di questo effetto, si può concludere che una tecnica efficace per sopprimere gli effetti di fringing del campo elettrico è l'utilizzo di un BOX sottile.

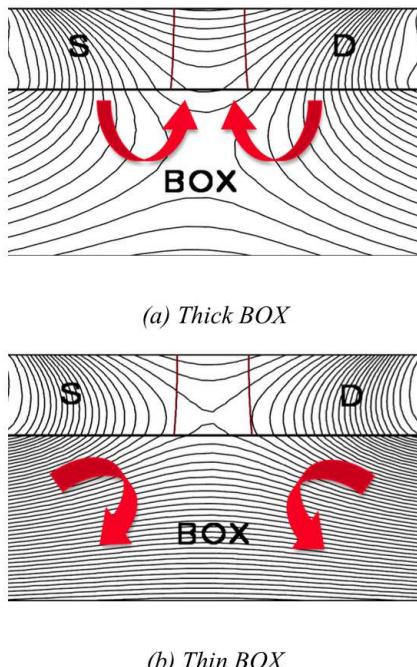


Figura 1.17: potenziale elettrostatico per un box spesso (60 nm) e uno sottile (15 nm) [11].

Le caratteristiche dei dispositivi FD-SOI CMOS sono influenzate, tra l'altro, anche dalle proprietà del substrato, in particolar modo dalla concentrazione di drogante N_B . A tal proposito è importante analizzare in che modo in un dispositivo FD-SOI sia possibile mantenere il controllo della tensione di soglia con il continuo scaling del dispositivo. Esistono principalmente due approcci per raggiungere la tensione di soglia desiderata; il primo consiste nel controllo della concentrazione di drogante N_B . Al fine di aumentare la tensione di soglia V_{TH} si dovrebbe aumentare il droggaggio di substrato, ma questo comporterebbe diverse limitazioni e conseguenze negative:

- la V_{TH} risulterebbe dipendente dalla variazione dello spessore del film di silicio;
- un'elevata concentrazione di drogante comporta un degrado della mobilità dei portatori e perdite ai bordi della giunzione a causa dell'effetto quantistico di *tunneling*;

- in dispositivi con body ultra sottile, la quantità di drogante richiesta per la desiderata V_{TH} non può essere praticamente raggiunta.

Per evitare i problemi derivanti dall'elevata densità di drogante dell'UTB, per il controllo della tensione di soglia può essere utilizzato un gate con funzione lavoro (ϕ_G) sintonizzabile, su un UTB non drogato [11]. Sebbene un canale SOI privo di drogante elimini i problemi relativi all'alto N_B , la fattibilità della soluzione descritta è fortemente soggetta alla disponibilità dei materiali necessari per la realizzazione dell'elettrodo di gate e dipende dalla facilità con cui questi possono essere integrati nel flusso di processo. Rimane comunque vero che il droggaggio relativamente basso rappresenta un vantaggio per lo scaling. L'uso, infatti, di un canale non drogato:

- riduce la dipendenza di V_{TH} dallo spessore t_{Si} ;
- riduce l'effetto di scattering da parte degli atomi droganti;
- riduce il campo elettrico trasversale;
- riduce le perdite dovute al *tunneling*;
- riduce l'effetto di RDF.

Per scaling sempre più aggressivi, per i quali i *non-classical* MOSFET sembrano rappresentare la soluzione più concreta, oltre allo spessore del film di silicio, si riduce anche lo spessore dell'ossido di gate, t_{ox} , con conseguente aumento delle perdite per *tunneling*. Una soluzione ampiamente utilizzata, che riduce le perdite di un ordine di grandezza, è l'impiego del nitruro insieme all'ossido di silicio, come si vede dalla figura 1.18.

La figura 1.19 mostra, invece, le diverse componenti della corrente di *tunneling*. Il *tunneling* degli elettroni provenienti dalla banda di valenza (EBV), in particolar modo, genera la corrente di substrato, che può essere trascurata nei dispositivi bulk-CMOS, ma non nei SOI MOSFET, in particolare nei PD SOI. Questa corrente infatti carica o scarica il body, cambiando così la tensione di soglia e pregiudicando il funzionamento del circuito.

Tuttavia, nei FD-SOI MOSFET la corrente di perdita è notevolmente ridotta in quanto, come visto, questi dispositivi sono caratterizzati da un body non drogato (o drogato leggermente) e la carica nella regione di svuotamento è pressoché nulla, il che comporta la riduzione del campo elettrico verticale nel canale.

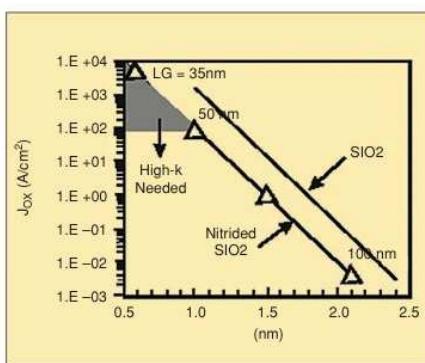


Figura 1.18: dipendenza della corrente di perdita dallo spessore di un ossido tradizionale (SiO_2) e da un ossido con nitruro [19].

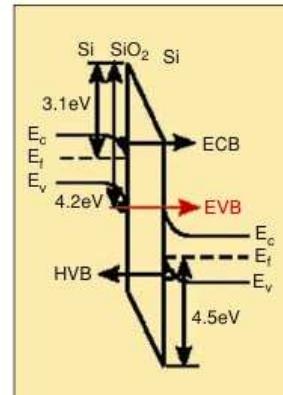


Figura 1.19: differenti componenti di *tunneling* in una struttura $Si/SiO_2/Si$ [19].

1.3.2 Tecnologie CMOS a gate multiplo

In aggiunta al FD-SOI MOSFET, il FinFET rappresenta un secondo candidato per i futuri dispositivi nanometrici. Come visto nei precedenti paragrafi, attualmente i due ostacoli principali allo scaling dei dispositivi CMOS sono le perdite di sottosoglia e del dielettrico di gate. Le strutture a gate multiplo, in particolar modo i *double gate (DG) FET* o i *triple gate (TG) FET*, possono essere utilizzate per far fronte a questi problemi poiché consentono un miglior controllo sugli effetti di canale corto.

L'architettura base di un DG MOSFET è caratterizzata dalla presenza di un secondo gate, opposto al primo (il gate tradizionale) come rappresentato in figura 1.20. Come mostrato nella parte destra della figura 1.20, quando la larghezza di canale del dispositivo è ridotta, il potenziale di drain comincia ad influenzare fortemente il potenziale di canale rendendo meno efficace il controllo da parte dell'elettrodo di gate. Questo effetto viene mitigato utilizzando un ossido di gate sottile e riducendo la regione di svuotamento al di sotto del canale. Tutto questo è stato discusso nei paragrafi precedenti, dove si è inoltre visto come la continua diminuzione dello spessore dell'ossido e della regione di svuotamento, di spessore X_D , comporti diversi problemi.

L'idea principale di un MOSFET a doppio gate è quella di controllare il canale in maniera efficiente scegliendo una larghezza di canale stretta e applicando un gate su entrambi i lati del canale stesso, schermandolo e permettendone

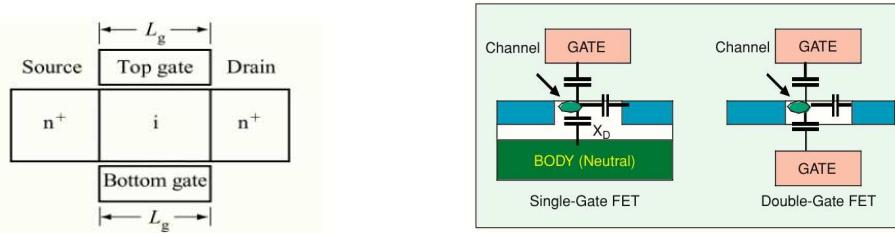


Figura 1.20: struttura base di un DG MOSFET (a sinistra), e vantaggi rispetto ad un FET Single-Gate (a destra) [20].

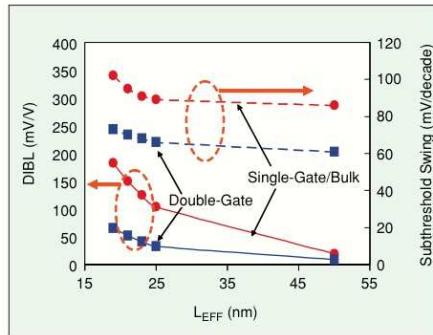


Figura 1.21: previsioni, ottenute mediante simulazioni fisiche di dispositivo, sulla variazione della pendenza di sottosoglia e del DIBL, in funzione della lunghezza di canale per un NFET DG e un NFET SG [20].

un miglior controllo. Questo riduce, inoltre, gli effetti di canale corto, in particolare l'effetto di DIBL. Si ha in aggiunta un aumento della pendenza di sottosoglia, in quanto il buon controllo elettrostatico effettuato da entrambi i lati del canale, riduce drasticamente l'influenza del campo di drain sul resto della struttura. In questo modo lo scaling di un dispositivo DG può procedere oltre i limiti dei CMOS convenzionali. La figura 1.21 mostra una previsione sulla pendenza di sottogola e sul DIBL per un dispositivo bulk-CMOS e uno double-gate, in funzione della lunghezza di canale effettiva L_{eff} .

Esistono tre tipologie principali di transistor a doppio gate, illustrate in figura 1.22:

- DG MOSFET planare, che è un'estensione diretta del processo CMOS planare con un secondo gate sepolto;
- DG MOSFET verticale, in cui il body (la pinna o fin) è verticale e source e drain sono collocati sulla sua parte superiore ed inferiore della pinna,

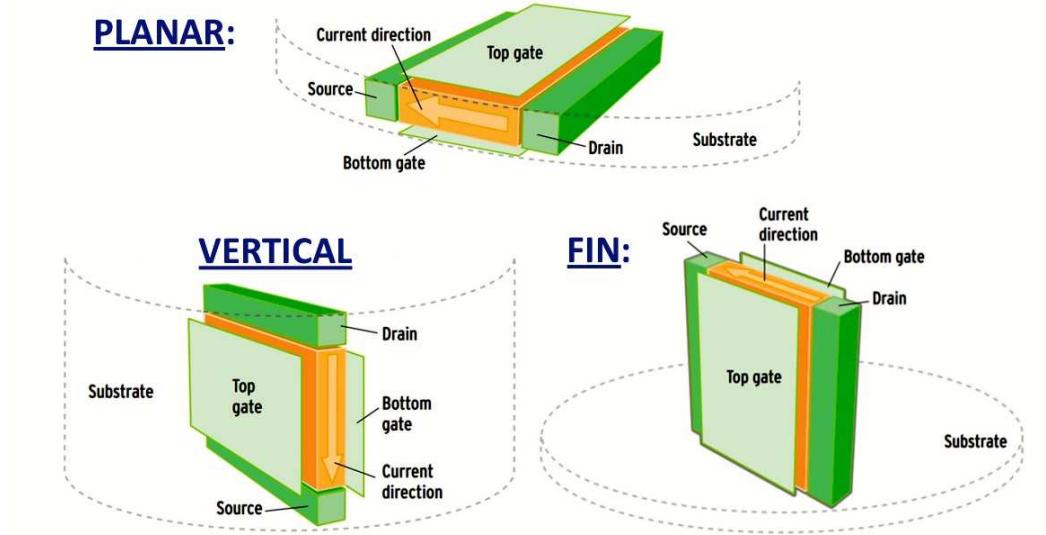


Figura 1.22: tipologie principali dei DG-MOSFET [21].

mentre il gate si trova sulle due facce laterali;

- FinFET, discusso nel paragrafo successivo, in cui il body è, come nel caso precedente, disposto in verticale e le regioni di source e drain sono poste alle sue estremità lungo la direzione orizzontale, come in un convenzionale FET planare.

Esistono, tuttavia, alcuni problemi che caratterizzano i dispositivi a doppio gate. Come indicato in figura 1.23 i principali sono:

1. la necessità di definire due gate con le stesse dimensioni;
2. l'autoallineamento delle regioni di drain e source sia per il gate superiore sia per quello inferiore;
3. l'allineamento dei due gate;
4. la necessità di fornire un collegamento tra i due gate mediante un percorso a bassa resistenza.

I primi tre aspetti sono strettamente correlati fra di loro e molto critici per i dispositivi a canale corto.

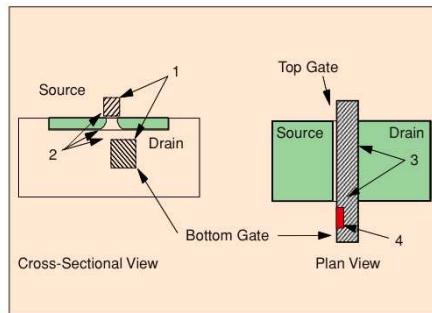


Figura 1.23: i quattro principali problemi di un MOSFET Double Gate [20].

Per i dispositivi DG MOSFET planari, questi requisiti risultano molto sfavorevoli in quanto il secondo gate è appunto sepolto sotto l'area attiva del MOS, a differenza dei FinFET, i cui gate sono facilmente accessibili.

1.3.3 Tecnologie FinFET

La figura 1.24 mostra, schematicamente, i passi relativi al processo produttivo di un FinFET, a confronto con quelli di un MOSFET-SOI, allo scopo di evidenziare le somiglianze tra i due processi. È possibile utilizzare come materiale di partenza un convenzionale wafer SOI, a meno dell'orientazione del piano cristallografico. La superficie del canale di un FinFET giace sul piano cristallografico (110) quando la pinna è orientata parallelamente o perpendicolarmente al flat o notch di un wafer standard (100). In questo caso la mobilità delle lacune è maggiore rispetto alla mobilità degli elettroni che risulta invece

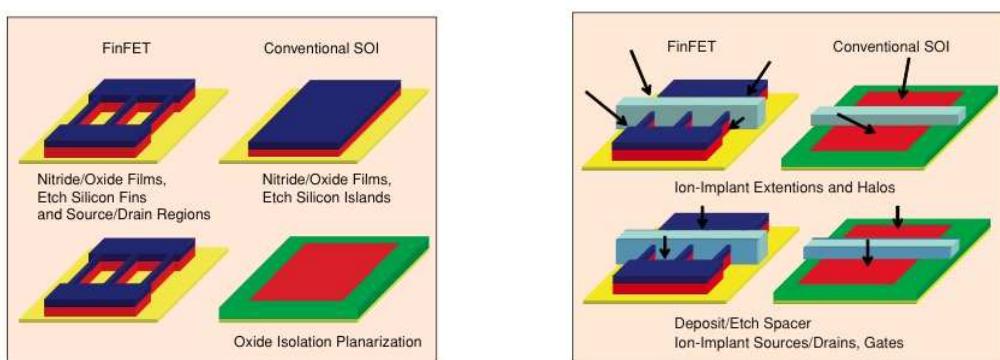


Figura 1.24: passi del processo FinFET [20].

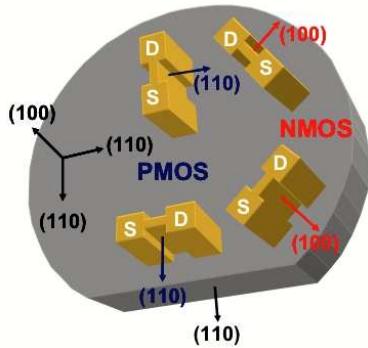


Figura 1.25: orientazione delle pinne per ottimizzare le prestazioni dei CMOS FinFET. I fin dei dispositivi PMOS giacciono sulla superficie (110), le pinne dei transistori NMOS sulla superficie (100) [11].

migliore quando la superficie del canale giace sul piano cristallografico (100). Di conseguenza, per massimizzare le prestazioni sia dei transistori NMOS che dei PMOS, la pinna dei PMOS è disposta in maniera parallela o perpendicolare rispetto al notch del wafer con orientazione (100) mentre la pinna degli NMOS è ruotata di un angolo di 45°, come mostrato in figura 1.25. In figura 1.26 vengono mostrati, in maniera schematica, i passi necessari per ottenere la pinna, che, di fatto, diventa il body del transistor. Tale processo, insieme a quello per la definizione delle regioni di source e drain, è molto simile al processo utilizzato per la realizzazione delle *trench isolation* in un convenzionale processo CMOS. Quindi il FinFET è, sostanzialmente, un MOSFET planare

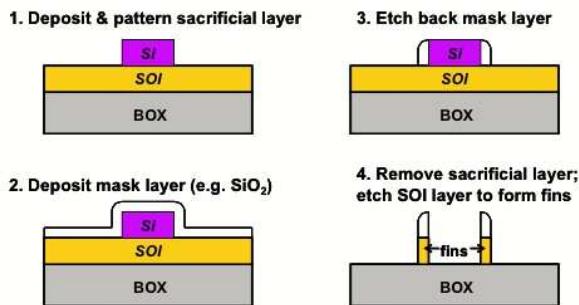


Figura 1.26: sequenza schematica che illustra il processo di formazione del body del dispositivo (la pinna). A questo scopo si utilizza generalmente la tecnica litografica denominata *spacer lithography technique* [22].

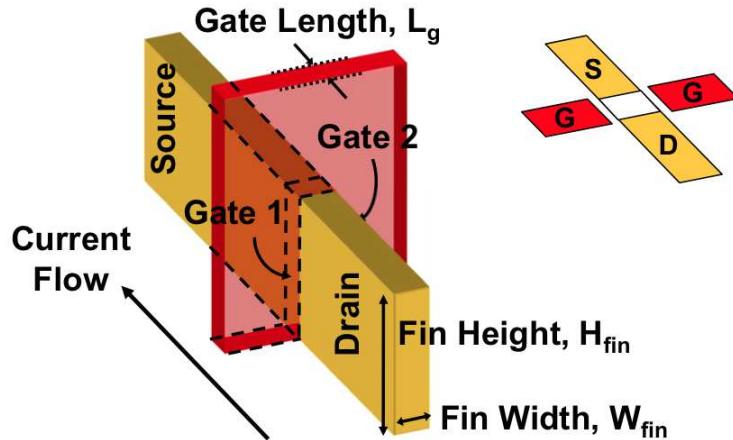


Figura 1.27: struttura di un FinFET [22].

orientato verticalmente, con il gate disposto tutto intorno all'*ultra thin body* (ovvero, alla pinna). Come mostrato in figura 1.27, la larghezza W effettiva del canale è due volte l'altezza della pinna più il suo spessore ($W = 2 \cdot H_{FIN} + T_{Si}$). La larghezza del body (o equivalentemente lo spessore della pinna) rappresenta ora la dimensione minima (al posto della L_g del processo planare). Si ha in questo modo l'introduzione di una nuova regola di scaling: i parametri che governano i requisiti di scalabilità e quindi gli effetti di canale corto sono le geometrie della pinna. La sfida alla miniaturizzazione si traduce nella realizzazione di una sottile *fin*, e il controllo del drogaggio del body, così come la necessità di un ossido estremamente sottile e con comportamento isolante il più possibile vicino a quello ideale, non risulta più essenziale.

Sperimentalmente si trova che lo spessore della *fin* (T_{Si}), necessario a rendere la corrente di leakage trascurabile soddisfa la seguente relazione [22]:

$$T_{Si} < \frac{2}{3} L_G, \quad (1.38)$$

dove L_G rappresenta la lunghezza di canale, in accordo con figura 1.27. Dall'espressione (1.38) si nota che il nuovo limite allo *scaling* è appunto lo spessore della pinna, e cioè la larghezza del body.

Diversi studi hanno dimostrato le eccellenti prestazioni di questi dispositivi. La figura 1.28, ad esempio, mostra le caratteristiche corrente-tensione di un CMOS FinFET con lunghezza di canale minima di 22 nm e tensione di alimentazione pari a 0.8 V, che evidenziano una buona immunità agli effetti di canale corto ($DIBL \approx 50$ mV, pendenza di sottosoglia $\approx 70 \frac{mV}{dec}$)

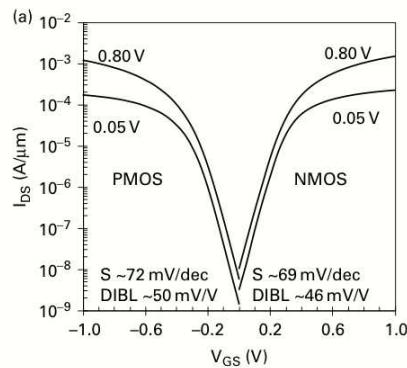


Figura 1.28: caratteristica $I - V$ per un dispositivo FinFET [11].

Nel processo schematizzato in figura 1.24 si è partiti da un substrato SOI. Questa tuttavia non è l'unica scelta: il FinFET può infatti essere realizzato anche su un substrato di silicio. Tale scelta è giustificata in modo particolare da ragioni di costo, dalla maggiore affidabilità del processo e dalla migliore conduzione termica del silicio. Esistono tuttavia degli svantaggi legati ai *bulk-Si FinFET* dovuti al processo di produzione e alle prestazioni elettriche di tali dispositivi, che fanno dei SOI FinFET una soluzione migliore, in particolar modo, quando la lunghezza di gate $L_g < 10$ nm.

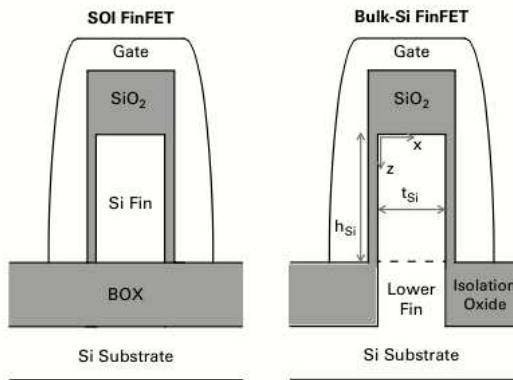


Figura 1.29: struttura base del SOI e del bulk-Si FinFET (sezione perpendicolare al canale) [11].

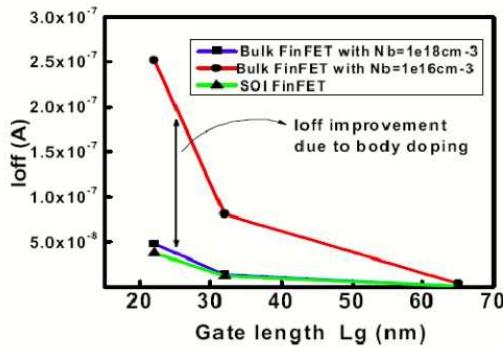


Figura 1.30: dipendenza della corrente di sottosoglia in funzione della lunghezza L_g , per un SOI FinFET e per un bulk FinFET con differenti drogaggi del *lower fin* [11].

La differenza principale tra i due FinFET che utilizzano i due differenti substrati, illustrati nella figura 1.29, risiede nella diversa formazione della pinna. I FinFET di tipo bulk richiedono che la parte inferiore della pinna (indicata in figura come *Lower Fin*) sia drogata pesantemente ($N_B > 10^{18} cm^{-3}$) per evitare il fenomeno di punch-through, e quindi per contenere la corrente di leakage, come mostrato in figura 1.30. Questo comporta una maggiore difficoltà nel raggiungere determinate prestazioni, in quanto il controllo della concentrazione di drogante N_B è cruciale, complicazione che in un SOI FinFET è assente grazie alla presenza del BOX sottostante.

Un altro problema che si presenta nei Bulk FinFET risiede nella complessità dei passi del processo di fabbricazione, principalmente a causa della criticità dell’impiantazione di drogante nel *lower fin* sopra discussa.

Capitolo 2

Caratterizzazione statica in dispositivi FinFET

In questo capitolo viene trattata la caratterizzazione statica di dispositivi NMOS e PMOS realizzati in tecnologia FinFET con lunghezza di canale minima pari a 14 nm . Dopo un breve cenno sui dispositivi sotto misura si discutono i principali parametri utilizzati per la descrizione del loro comportamento in termini statici e di segnale e si analizzano i risultati ottenuti.

2.1 Descrizione dei dispositivi sotto misura

I dispositivi studiati sono FinFET a canale N e a canale P, caratterizzati da differenti dimensioni di gate. I DUT sono stati resi disponibili sia come *naked dice*, successivamente connessi tramite *wire bonding* ad un pakage LCC44, sia direttamente nel pakage TQFP 176L. Per rendere possibile la caratterizzazione in termini di densità spettrale di rumore, descritta nel capitolo successivo, i transistor sono stati progettati con una larghezza di canale non inferiore a $100\text{ }\mu\text{m}$. Ogni DUT è costituito dalla connessione in parallelo di un numero n_d di dispositivi elementari, ognuno dei quali consiste in 8 *fin*, ciascuna con larghezza effettiva di canale di 74 nm. Di conseguenza la larghezza di gate W di ogni dispositivo nella struttura di test è un multiplo intero di $8 \cdot 74\text{ nm}$, mentre la lunghezza del gate L assume valori nell'intervallo 14 nm - 100 nm. La tabella 2.1 fornisce una lista dei transistor disponibili, specificando la larghezza e la lunghezza di gate, il numero n_d di dispositivi elementari ed il tipo di pakage. La tensione di alimentazione nominale è pari a $V_{DD} = 0.8\text{ V}$.

NMOS				
W [μm]	L [nm]	n _d	available in	
			TQFP 176L pkg	LCC44 pkg
100	14	170	yes	yes
	18		no	yes
	80		no	yes
	100		no	yes
200	14	338	yes	yes
	18		no	yes
	80		yes	no
	100		no	yes
600	14	1014	yes	no
	18		no	yes
	80		yes	yes
	100		no	yes

PMOS				
W [μm]	L [nm]	n _d	available in	
			TQFP 176L pkg	LCC44 pkg
100	14	170	yes	yes
	18		no	yes
	80		no	yes
	100		no	yes
200	14	338	yes	yes
	18		no	yes
	80		yes	no
	100		no	yes
600	14	1014	yes	no
	18		no	yes
	80		yes	yes
	100		no	yes

Tabella 2.1: dispositivi FinFET a canale N e P disponibili per la caratterizzazione.

La figura 2.1 mostra la piedinatura del pakage LCC44 con la relativa *pin list*, mentre la figura 2.2, riporta i *bonding pads*¹ del chip contenente le strutture di test.

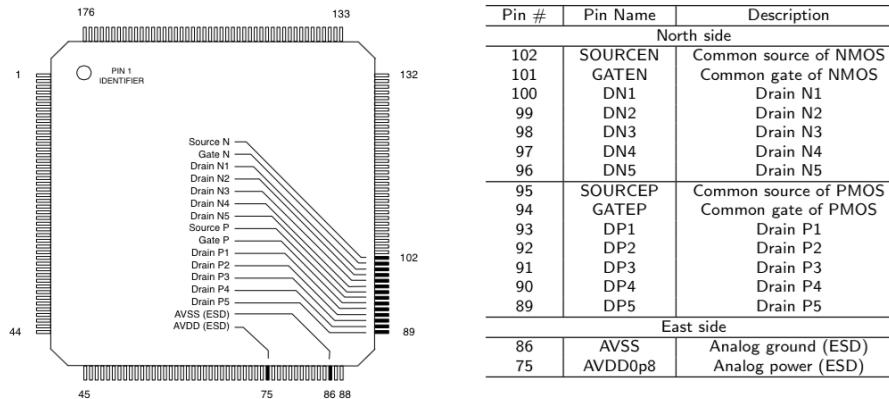


Figura 2.1: pin list e diagramma di bonding del pakage TQFP 176L.

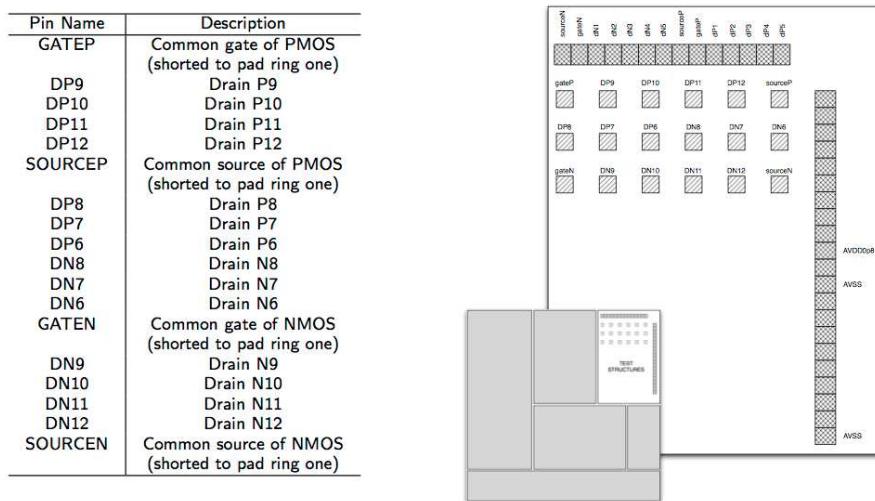


Figura 2.2: bonding pads del chip connessi al pakage LCC44.

¹I pads hanno dimensione pari a: 48 μm x 60 μm e sono distanziati gli uni dagli altri di 50 μm x 50 μm .

2.2 Caratteristiche statiche

L'analisi delle caratteristiche statiche di dispositivi a semiconduttore permette, ad esempio, la generazione delle curve $I - V$ e l'estrapolazione di parametri che, da un lato, forniscono informazioni circa la qualità del processo produttivo, dall'altro consentono di effettuare previsioni teoriche del comportamento dei dispositivi dal punto di vista delle caratteristiche di rumore. Per la caratterizzazione statica di dispositivi a semiconduttore è possibile effettuare misure (generalmente quasi statiche) in maniera semplice ed affidabile mediante strumenti programmabili, quali gli analizzatori di parametri di semiconduttore. Essi integrano un certo numero di SMU (Source-Measurement Unit) opportunamente programmate e sincronizzate all'interno di un *mainframe*, che ne consente la gestione mediante pannello di controllo e display sullo strumento stesso o tramite interfaccia software su calcolatore.

In questo lavoro le misure statiche sono state condotte utilizzando lo strumento *HP4145B Semiconductor Parameter Analyzer*. Tale strumento, dotato di porta GPIB, è interfacciato con un calcolatore mediante un programma sviluppato in ambiente LabView, che permette di associare i canali dello strumento di misura ai terminali del DUT e di scegliere il tipo di sweep (lineare o logaritmico), l'intervallo di variazione di tensioni e correnti, i valori di *compliance* (il massimo valore di tensione applicabile ad un nodo o di corrente che possa fluire in un terminale del DUT) ed infine le variabili da visualizzare e la loro memorizzazione.

Su ciascun FinFET sono state effettuate misure della corrente di drain (I_D) e di gate (I_G) in funzione di:

- V_{GS} (V_{SG} per i PMOS) che varia da 0 a 0.8 V, con V_{DS} (V_{SD}) come parametro tra 0 e 0.8 V con step di 0.2 V (0, 0.2, 0.4, 0.6, 0.8 V);
- V_{DS} (V_{SD}) che varia da 0 a 0.8 V, con V_{GS} (V_{SG}) come parametro tra 0.2 V e 0.8 V con step di 0.2V (0.2, 0.4, 0.6, 0.8 V).

Per evitare il possibile danneggiamento dei dispositivi, la *compliance* imposta sulla corrente di drain è 100 mA, mentre sulla corrente di gate è stata fissata a 16 mA. Il limite relativamente alto imposto sulla corrente di gate è giustificato dal suo elevato valore misurato in fase di caratterizzazione. La causa di ciò è da attribuirsi a contributi di diversa natura provenienti dai differenti dispositivi presenti nella struttura di test che hanno i terminali di gate e di source in comune. Gli andamenti delle curve $I_D - V_{DS}$ e $I_D - V_{GS}$, sono mostrati, a titolo di esempio, nelle figure dalla 2.3 alla 2.10.

Dalle caratteristiche statiche $I_D - V_{GS}$ sono stati ricavati gli andamenti della

transconduttanza g_m al variare della V_{GS} . Alcune delle curve ricavate sono mostrate nelle figure dalla 2.11 alla 2.14.

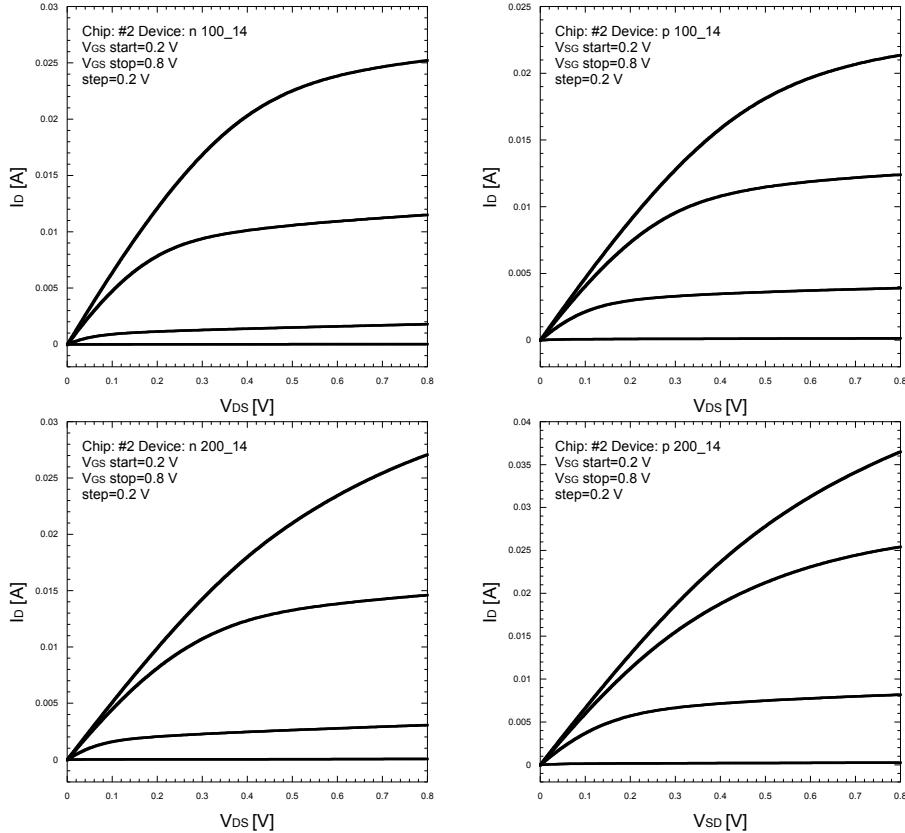


Figura 2.3: corrente di drain I_D in funzione della tensione drain-source V_{DS} con V_{GS} come parametro per dispositivi a canale N (a sinistra) e a canale P (a destra) con lunghezza di canale $L = 14$ nm.

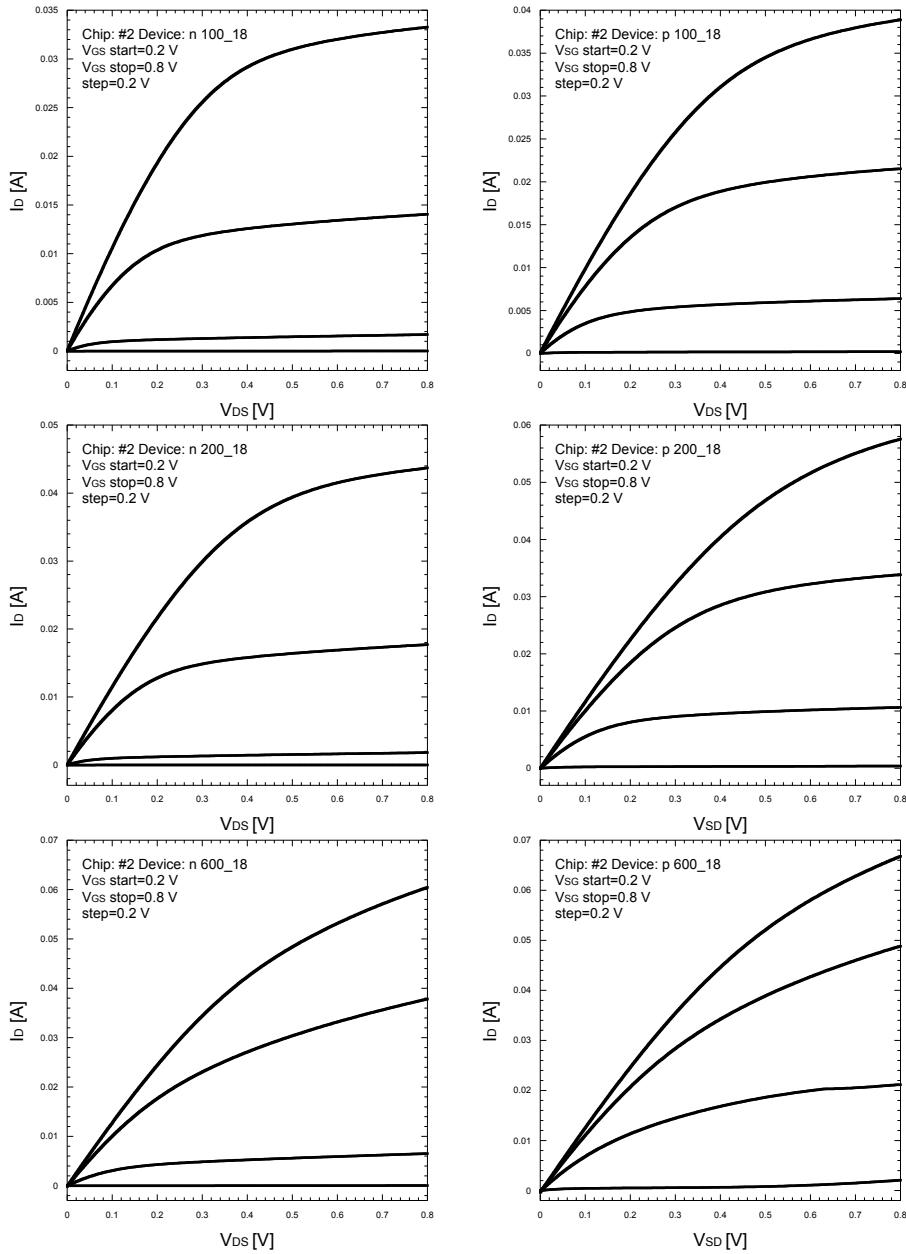


Figura 2.4: corrente di drain I_D in funzione della tensione drain-source V_{DS} con V_{GS} come parametro per dispositivi a canale N (a sinistra) e a canale P (a destra) con lunghezza di canale $L = 18$ nm.

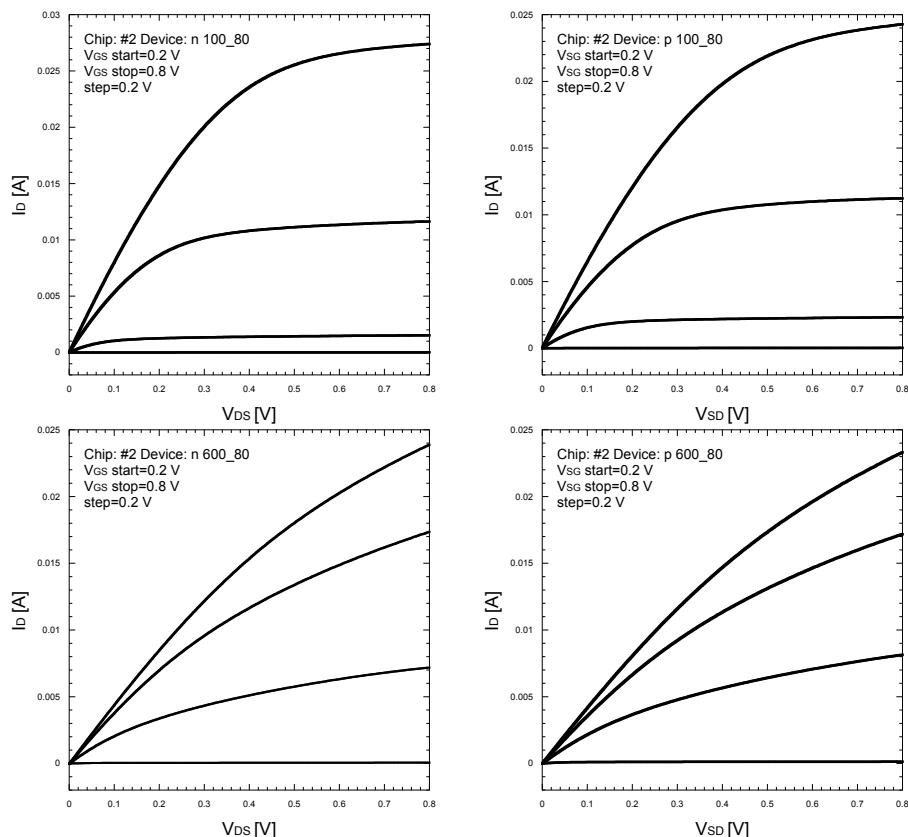


Figura 2.5: corrente di drain I_D in funzione della tensione drain-source V_{DS} con V_{GS} come parametro per dispositivi a canale N (a sinistra) e a canale P (a destra) con lunghezza di canale $L = 80$ nm.

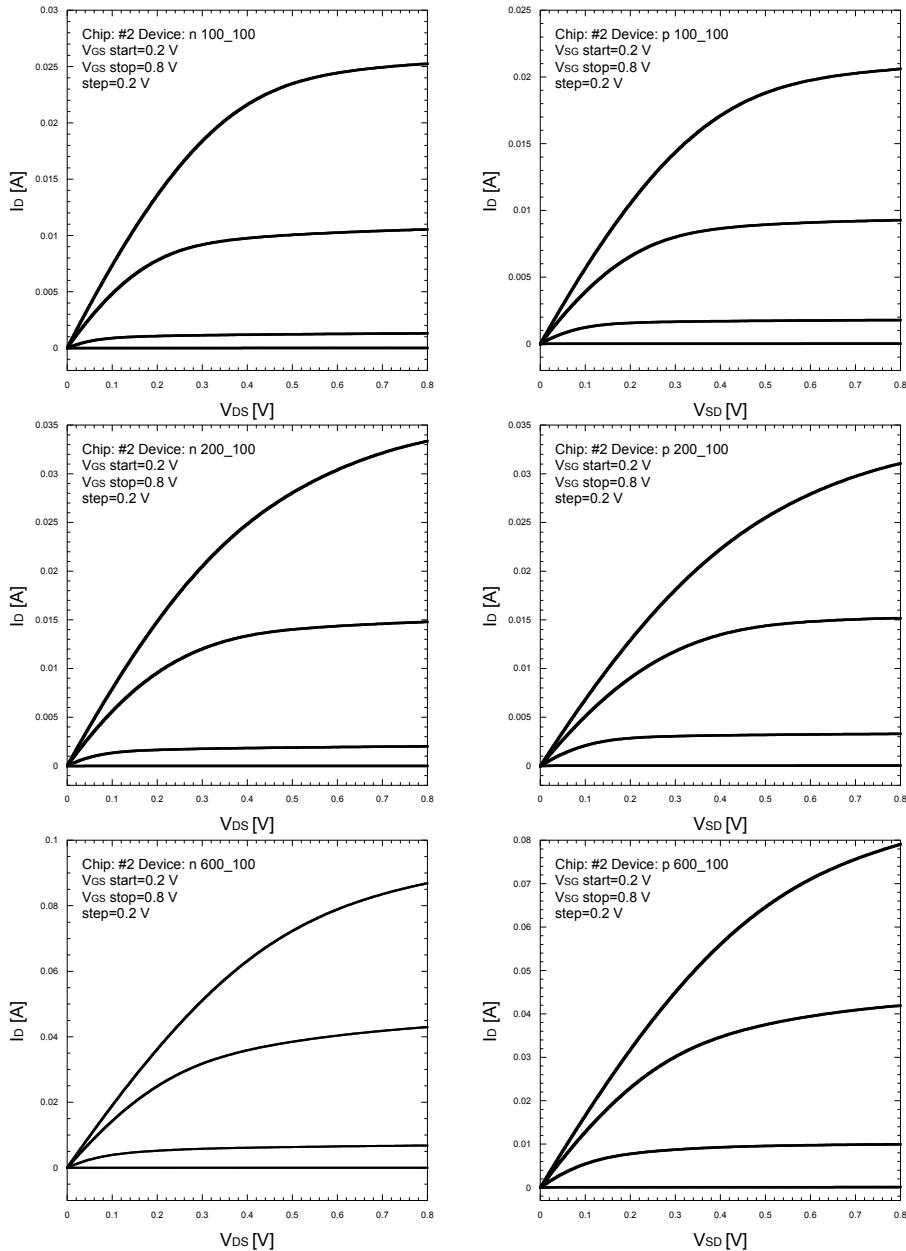


Figura 2.6: corrente di drain I_D in funzione della tensione drain-source V_{DS} con V_{GS} come parametro per dispositivi a canale N (a sinistra) e a canale P (a destra) con lunghezza di canale $L = 100$ nm.

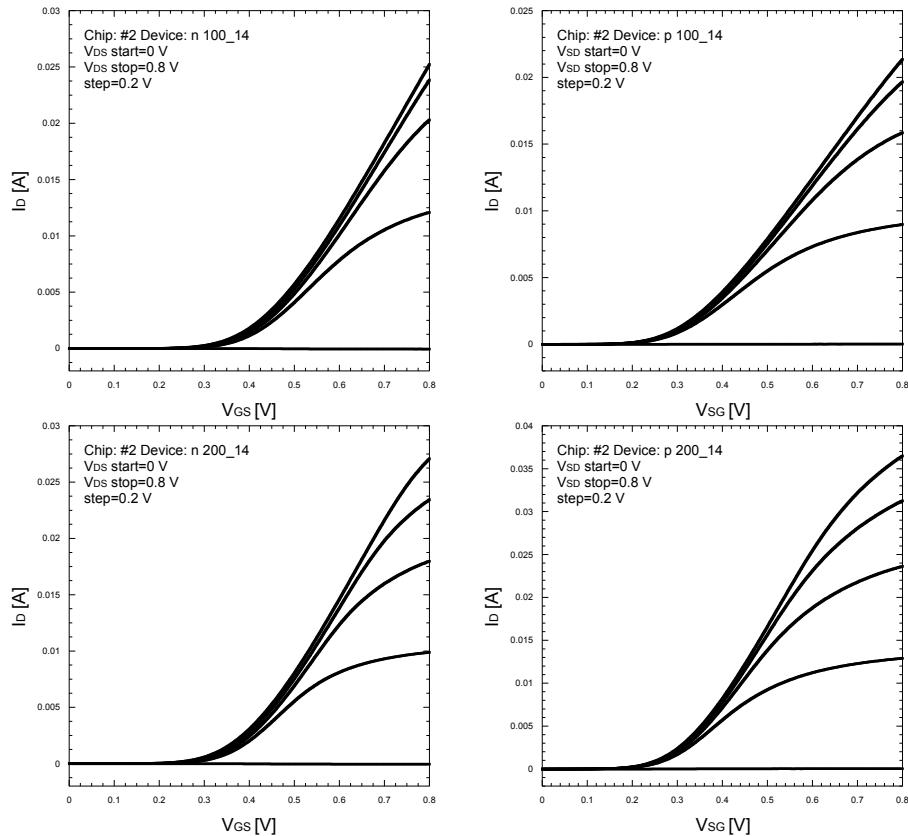


Figura 2.7: corrente di drain I_D in funzione della tensione gate-source V_{GS} con V_{DS} come parametro per dispositivi a canale N (a sinistra) e a canale P (a destra) con lunghezza di canale $L = 14$ nm.

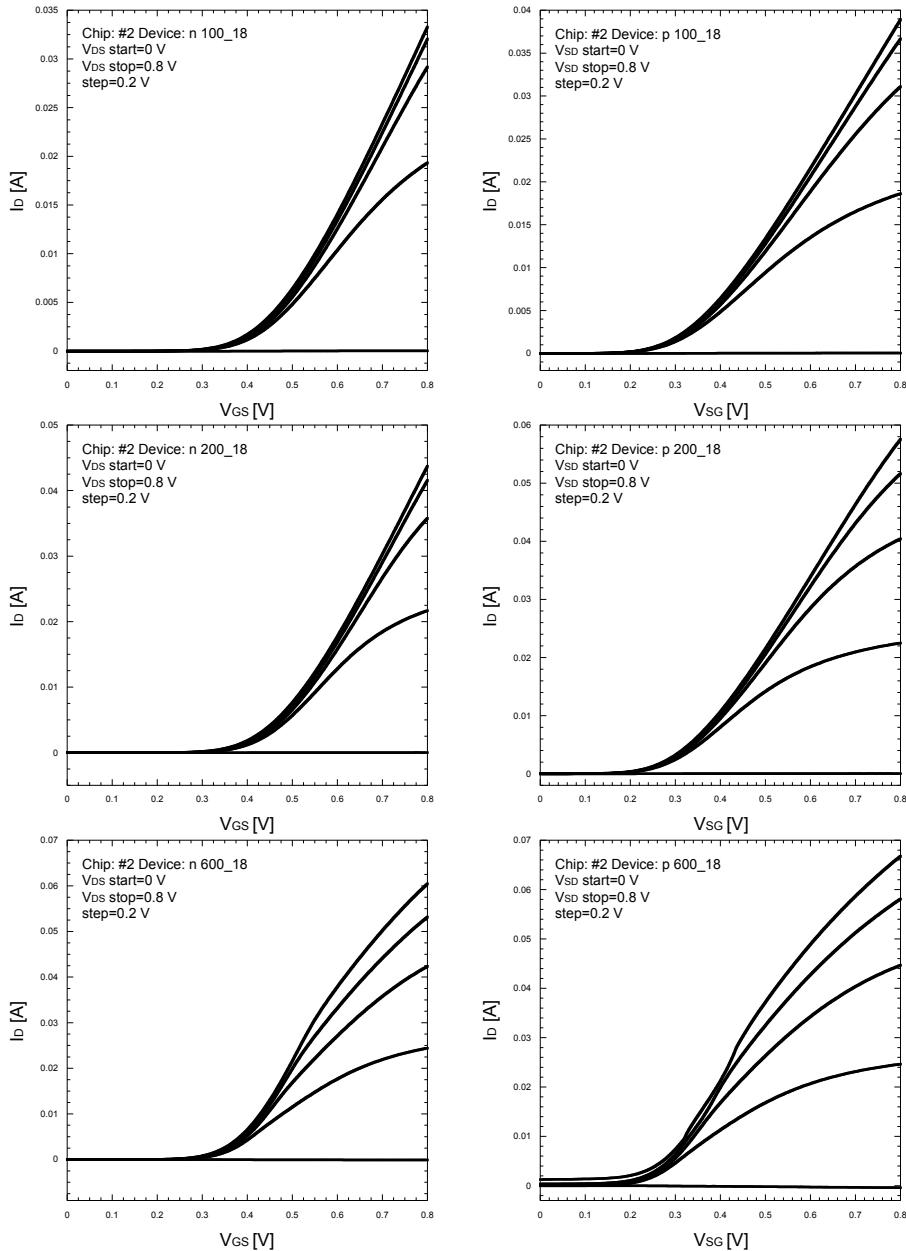


Figura 2.8: corrente di drain I_D in funzione della tensione gate-source V_{GS} con V_{DS} come parametro per dispositivi a canale N (a sinistra) e a canale P (a destra) con lunghezza di canale $L = 18$ nm.

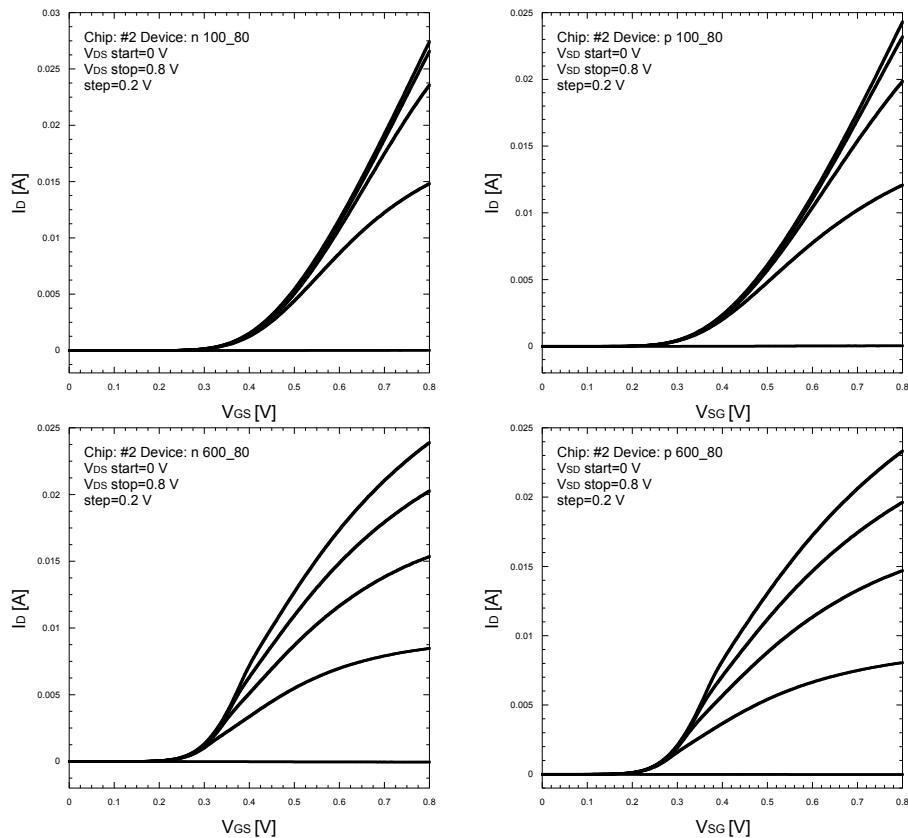


Figura 2.9: corrente di drain I_D in funzione della tensione gate-source V_{GS} con V_{DS} come parametro per dispositivi a canale N (a sinistra) e a canale P (a destra) con lunghezza di canale $L = 80$ nm.

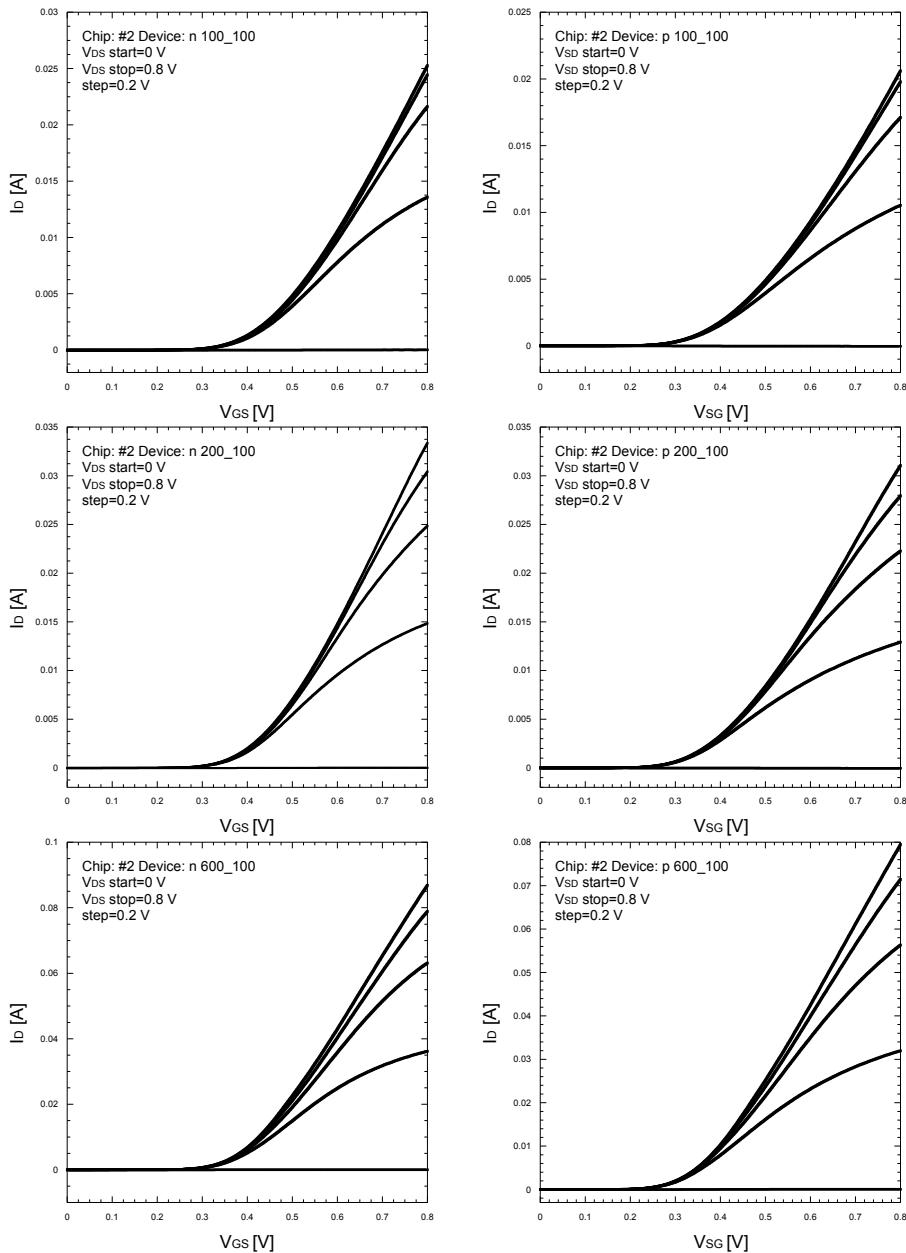


Figura 2.10: corrente di drain I_D in funzione della tensione gate-source V_{GS} con V_{DS} come parametro per dispositivi a canale N (a sinistra) e a canale P (a destra) con lunghezza di canale $L = 100$ nm.

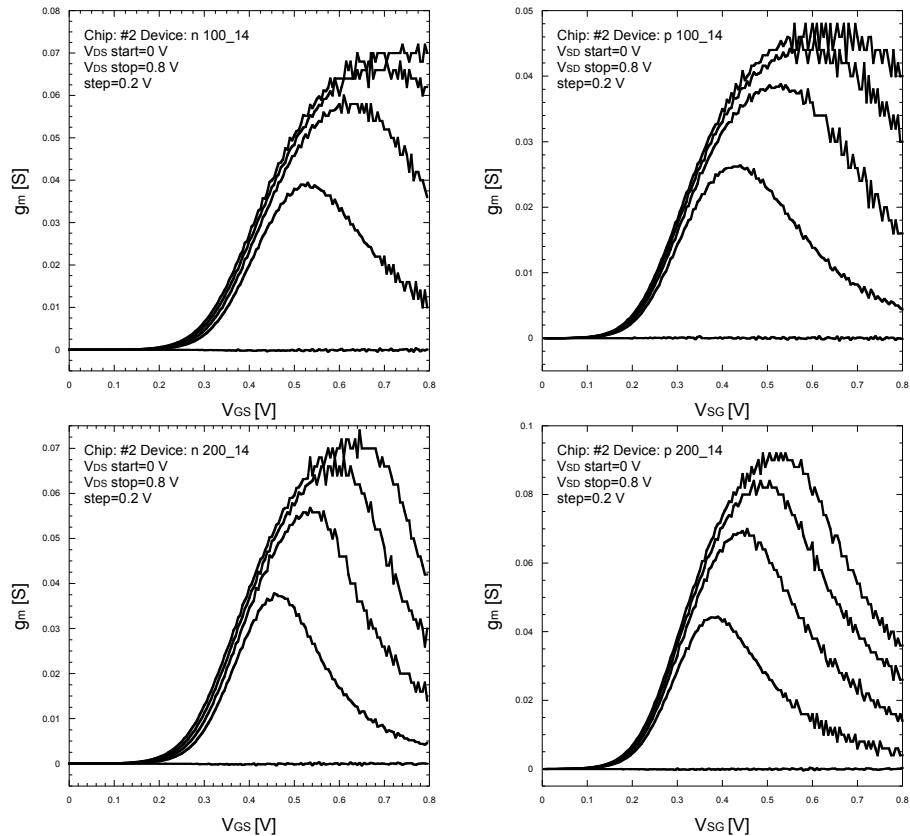


Figura 2.11: transconduttanza g_m in funzione della tensione V_{GS} con V_{DS} come parametro per dispositivi a canale N (a sinistra) e a canale P (a destra) con lunghezza di canale $L = 14$ nm.

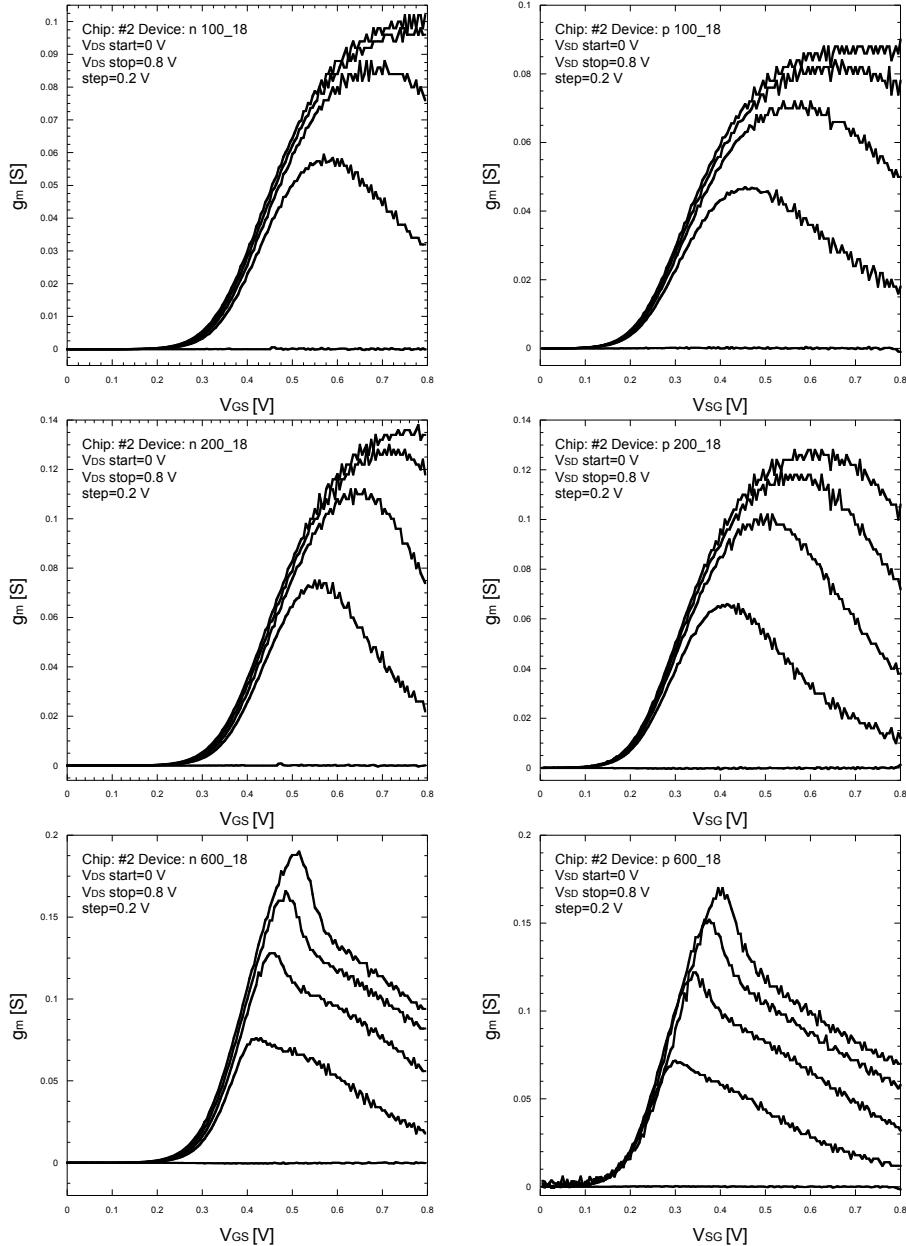


Figura 2.12: transconduttanza g_m in funzione della tensione V_{GS} con V_{DS} come parametro per dispositivi a canale N (a sinistra) e a canale P (a destra) con lunghezza di canale $L = 18$ nm.

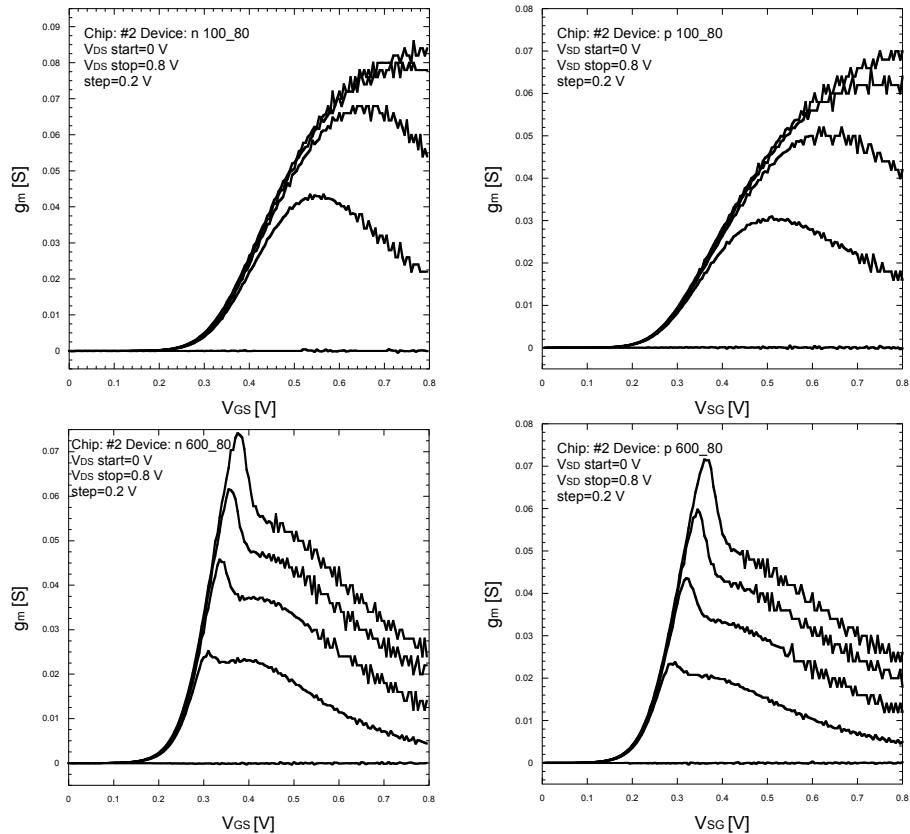


Figura 2.13: transconduttanza g_m in funzione della tensione V_{GS} con V_{DS} come parametro per dispositivi a canale N (a sinistra) e a canale P (a destra) con lunghezza di canale $L = 80$ nm.

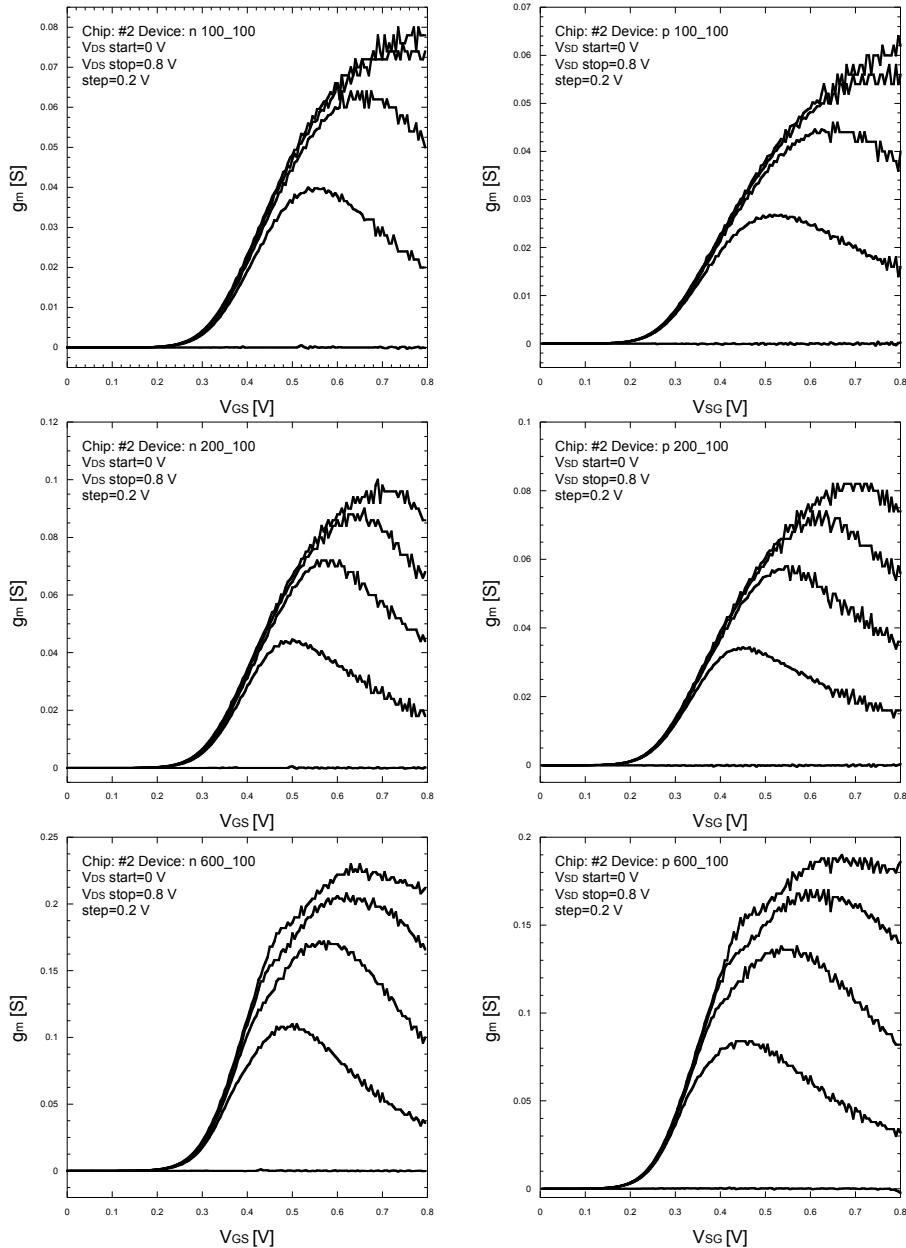


Figura 2.14: transconduttanza g_m in funzione della tensione V_{GS} con V_{DS} come parametro per dispositivi a canale N (a sinistra) e a canale P (a destra) con lunghezza di canale $L = 100$ nm.

2.2.1 Corrente di drain caratteristica normalizzata

In molte applicazioni, dove una bassa dissipazione di potenza rappresenta un requisito indispensabile, i dispositivi MOSFET si trovano ad operare in regione di debole o moderata inversione. Lo studio del loro comportamento in questa regione di lavoro può trarre vantaggio dalla definizione di alcuni concetti e parametri quali l'efficienza di transconduttanza, definita come il rapporto tra la transconduttanza g_m del transistor e la corrente di drain I_D , ed il coefficiente di inversione.

In debole inversione, la transconduttanza g_m può essere espressa come:

$$g_m = \frac{I_D}{n_{sub}V_t}, \quad (2.1)$$

dove $V_t = \frac{k_B T}{q}$ è la tensione termica ed n_{sub} è un coefficiente che si ricava sperimentalmente dalle caratteristiche $I_D - V_{GS}$ in regione di sottosoglia e che dipende dalla pendenza di sottosoglia S , trattata più avanti nel capitolo. Dalla (2.1) si ricava in maniera immediata, sempre in debole inversione (w.i., weak inversion), dove essa ha valore, che il rapporto g_m/I_D , ovvero l'efficienza di transconduttanza, è costante e pari a:

$$\left(\frac{g_m}{I_D}\right)_{w.i.} = \frac{1}{n_{sub}V_t}. \quad (2.2)$$

In forte inversione (s.i., strong inversion), invece, la transconduttanza è proporzionale alla radice quadrata della corrente di drain:

$$g_m = \sqrt{2\frac{\mu C_{ox} \frac{W}{L}}{n_{sub}} I_D}, \quad (2.3)$$

da cui si ottiene l'espressione dell'efficienza di transconduttanza valida in forte inversione:

$$\left(\frac{g_m}{I_D}\right)_{s.i.} = \sqrt{2\frac{\mu C_{ox} \frac{W}{L}}{n_{sub}} \frac{1}{I_D}}. \quad (2.4)$$

È possibile definire una corrente caratteristica di drain normalizzata I_Z^* che separa la regione di debole inversione da quella di forte inversione, data da [23]:

$$I_Z^* = 2\mu C_{ox} n_{sub} V_t^2, \quad (2.5)$$

dove μ è la mobilità dei portatori nel canale. La figura 2.15 a) mostra il rapporto g_m/I_D in funzione della corrente di drain normalizzata $I_D \cdot L/W$, che provvede a fornire un esempio del comportamento tipico dei dispositivi sotto

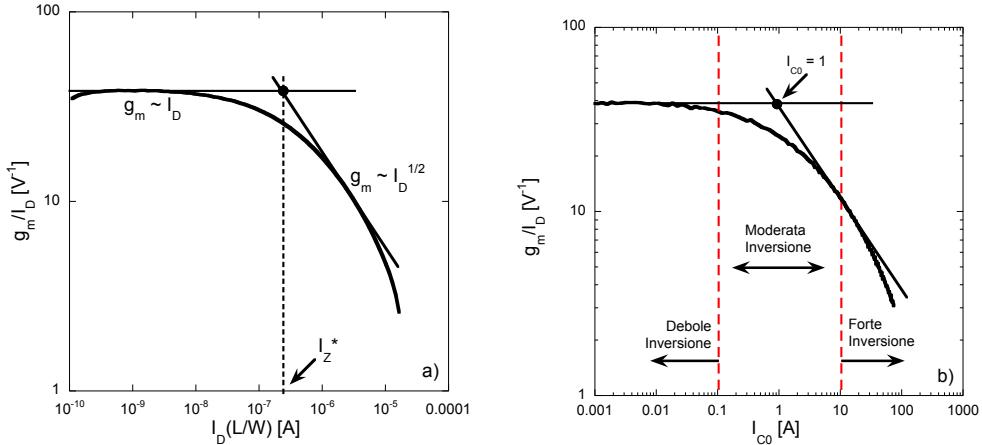


Figura 2.15: efficienza di transconduttanza, ricavata da misure sui FinFET, in funzione della a) corrente normalizzata di drain e b) del coefficiente di inversione I_{C0} .

misura. Per bassi valori della corrente di drain normalizzata, la transconduttanza è proporzionale alla corrente di drain I_D e, come detto, l'efficienza di transconduttanza è costante e tende ad approssimarsi ad un asintoto orizzontale, il cui valore può essere utilizzato per il calcolo del coefficiente n_{sub} mediante l'espressione (2.2). Per alti valori della $I_D \cdot (W/L)$, il DUT entra in regione di forte inversione, dove la pendenza del rapporto g_m/I_D , su scala logaritmica, diventa pari a $-1/2$, in quanto, come detto, in quella regione la transconduttanza è proporzionale alla radice quadrata della corrente di drain. La corrente caratteristica normalizzata I_Z^* è definita dunque come il valore di corrente normalizzata corrispondente all'intersezione tra la tangente alla curva in forte inversione e l'asintoto orizzontale alla curva in debole inversione. Si nota, infine, sempre dalla figura 2.15 che, per alti valori di I_D , la pendenza della curva g_m/I_D diventa più ripida. Questo fatto è determinato dalla saturazione della velocità (v.s., *velocity saturation*) dei portatori. Considerando, infatti, l'efficienza di trasconduttanza nel caso della saturazione di velocità, si ha [24]:

$$\left(\frac{g_m}{I_D}\right)_{v.s.} = \frac{WC_{ox}v_{sat}}{I_D}. \quad (2.6)$$

La corrente caratteristica normalizzata, come già detto, viene in genere utilizzata come valore di riferimento per separare la regione di funzionamento in debole inversione da quella in forte inversione del dispositivo. I_Z^* viene collocata

ta nella regione cosiddetta di moderata inversione che si estende per definizione una decade al sopra ed una decade al di sotto di essa. Queste considerazioni sulla regione operativa del dispositivo e sulla relazione con la corrente I_Z^* portano al concetto di coefficiente di inversione, che serve appunto per quantificare il livello di inversione del canale. Il coefficiente di inversione, I_{C0} , viene definito come [25]:

$$I_{C0} = \frac{I_D}{2n_{sub}\mu C_{ox}(\frac{L}{W})V_t^2}. \quad (2.7)$$

Una forma semplice per il coefficiente di inversione può essere trovata utilizzando, nella (2.7) l'espressione di I_Z^* . Dalla (2.5) si ottiene:

$$I_{C0} = \frac{I_D}{I_Z^*} \left(\frac{L}{W} \right). \quad (2.8)$$

Con riferimento alla figura 2.15 b), al centro della moderata inversione, vale a dire quando $I_D \cdot L/W = I_Z^*$, dalla (2.8) si ricava che il coefficiente di inversione è pari ad uno. Sotto l'ipotesi che la regione di moderata inversione, come già detto, si estenda una decade prima ed una decade dopo I_Z^* [25], si può assumere che il dispositivo entri in debole inversione quando I_{C0} è minore di 0.1 ed in forte inversione quanto I_{C0} è maggiore di 10. In accordo con la (2.5), I_Z^* risulta maggiore nei dispositivi NMOS rispetto ai PMOS, a causa della mobilità superiore degli elettroni. I risultati sperimentali ottenuti mostrano, tuttavia, una corrente di drain caratteristica di poco superiore nei FinFET a canale P rispetto ai dispositivi a canale N. Una possibile interpretazione circa questo risultato risiede nel forte impatto che l'orientazione della pinna, rispetto agli assi cristallografici del *wafer*, ha sulla mobilità dei portatori nei dispositivi MuGETs (Multiple Gate FETs) [26]. Come già detto nel capitolo 1, se la pinna del dispositivo giace sul piano cristallografico (110) la mobilità delle lacune è aumentata mentre quella degli elettroni viene diminuita. Tenendo conto della (2.5), il fatto che il valore della corrente caratteristica normalizzata sia simile nei dispositivi a canale P ed in quelli a canale N sembra indicare che la mobilità sia pure simile per lacune ed elettroni. Questo risultato si potrebbe giustificare col fatto che, nel processo produttivo dei FinFET studiati, le pinne sono state disposte parallelamente o perpendicolarmente al *flat* del *wafer* favorendo, in questo modo, la mobilità delle lacune a scapito di quella degli elettroni.

La figura 2.16 mostra il metodo di estrazione della corrente caratteristica normalizzata di drain, mettendo a confronto un PMOS ed un NMOS di uguali dimensioni. Dalla figura 2.17 alla 2.19 si presenta, a titolo esemplificativo, l'efficienza di transconduttanza in funzione della corrente di drain normalizzata per alcuni tra i dispositivi analizzati.

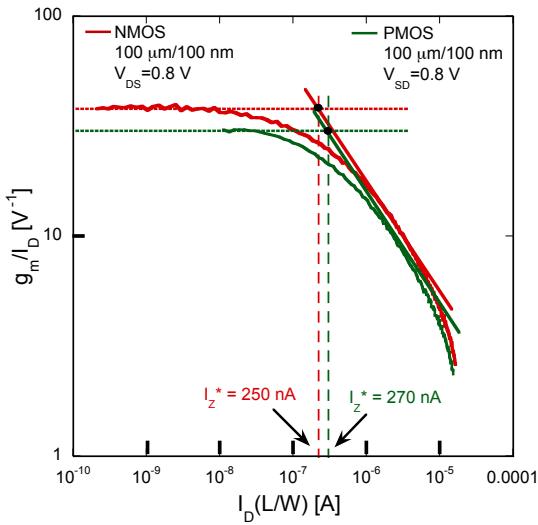


Figura 2.16: estrazione della corrente caratteristica di drain normalizzata I_Z^* per un FinFET a canale N ed un FinFET a canale P con uguali dimensioni di gate.

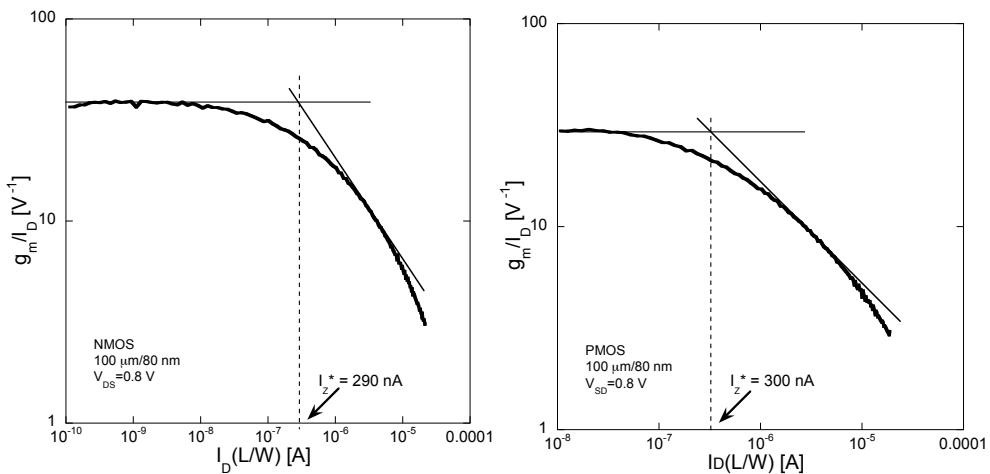


Figura 2.17: estrazione della corrente caratteristica I_Z^* per un FinFET a canale N (a sinistra) e per uno a canale P (a destra) con $W/L = 100 \mu\text{m}/80 \text{ nm}$.

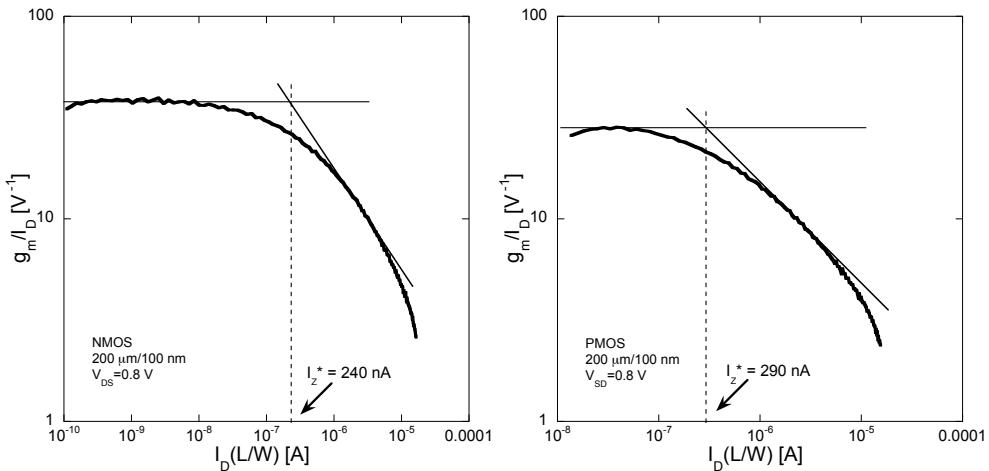


Figura 2.18: estrazione della corrente caratteristica I_Z^* per un FinFET a canale N (a sinistra) e per uno a canale P (a destra) con $W/L = 200 \mu\text{m}/100 \text{ nm}$.

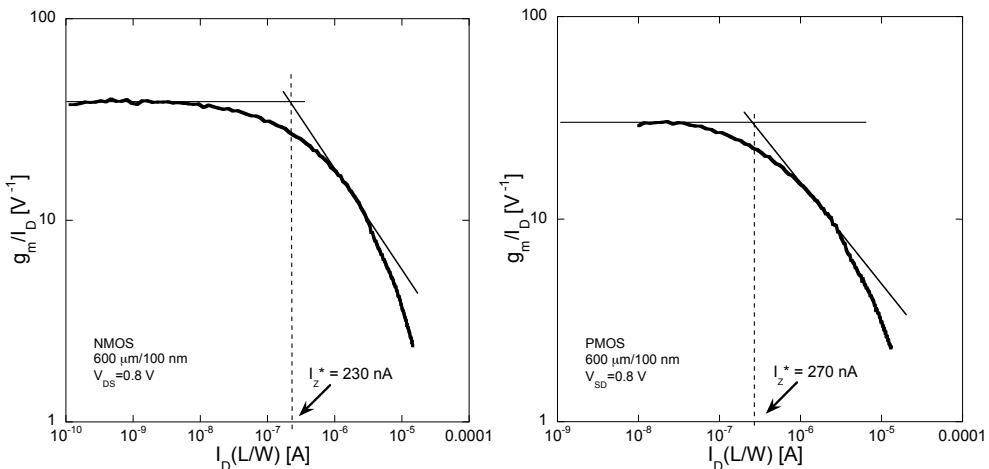


Figura 2.19: estrazione della corrente caratteristica I_Z^* per un FinFET a canale N (a sinistra) e per uno a canale P (a destra) con $W/L = 600 \mu\text{m}/100 \text{ nm}$.

	NMOS		PMOS	
	n_{sub}	$I_Z^* \text{ [nA]}$	n_{sub}	$I_Z^* \text{ [nA]}$
14 nm	1.1	250	1.3	270
65 nm	1.25	490	1.25	150
90 nm	1.2	760	1.2	200
130 nm	1.2	600	1.2	150

Tabella 2.2: coefficiente n_{sub} e corrente di drain caratteristica normalizzata I_Z^* estratta dai dispositivi sotto misura e da dispositivi appartenenti a differenti tecnologie bulk CMOS.

La tabella 2.2 riassume i valori estratti del coefficiente n_{sub} e della corrente di drain caratteristica I_Z^* per la tecnologia studiata. Nell'estrazione dei valori mostrati non si sono considerati i dispositivi con lunghezza minima di canale. Tale scelta è giustificata dal fatto che la lunghezza di gate effettiva è differente da quella nominale, specialmente nei dispositivi caratterizzati dalla lunghezza minima di canale consentita dalla tecnologia. Di conseguenza il valore estratto di I_Z^* e quello del coefficiente n_{sub} risentono maggiormente di questo effetto. A fini comparativi in tabella sono, inoltre, inclusi i valori caratteristici di alcune tecnologie bulk CMOS.

2.2.2 Guadagno di tensione intrinseco

Il guadagno di tensione intrinseco, A_{Vi} , è definito come il guadagno di tensione di piccolo segnale tra gate e drain a bassa frequenza con il source connesso a massa ed il drain connesso ad un generatore di corrente ideale (e quindi ad una resistenza infinita). Poiché A_{Vi} è il massimo guadagno ottenibile per un singolo transistor, esso rappresenta un figura di merito molto utilizzata per comprendere l'impatto dello scaling sulle prestazioni dei circuiti analogici.

Il guadagno di tensione intrinseco è uguale al rapporto tra la trasconduttanza di canale, g_m , e la conduttanza drain-source g_{ds} :

$$A_{Vi} = \frac{g_m}{g_{ds}}. \quad (2.9)$$

La transconduttanza di canale è data da $g_m = \partial I_D / \partial V_{GS}$, con le tensioni drain-source, V_{DS} , e source-body, V_{SB} , mantenute costanti e può essere espressa dalla seguente equazione, valida in saturazione ed in qualunque regione di funzionamento, dalla debole alla forte inversione [27]:

$$g_m = \frac{I_D}{n_{sub}V_t} \frac{2}{1 + \sqrt{1 + 4I_{C0}}}. \quad (2.10)$$

La conduttanza di uscita, g_{ds} , descrive la variazione della corrente di drain a fronte di un cambiamento della tensione V_{DS} , con V_{GS} e V_{SB} costanti, ed è data da $g_{ds} = \partial I_D / \partial V_{DS}$.

In questo lavoro di tesi il guadagno intrinseco A_{Vi} è studiato in funzione del livello di inversione I_{C0} , come mostrato in figura 2.20 per i dispositivi a canale N con differenti geometrie di gate. Il guadagno, come atteso, risulta massimo in debole inversione, dove è anche indipendente dalla corrente di drain. Infatti in questa regione operativa [28]:

$$g_m = \frac{I_D}{n_{sub} \frac{K_B T}{q}}; \quad (2.11)$$

$$g_{ds} = \lambda_{wi} I_D, \quad (2.12)$$

dove λ_{wi} è il fattore di modulazione della lunghezza di canale in debole inversione, inversamente proporzionale alla lunghezza L_G e K_B è la costante di Boltzmann. Dalla (2.11) si nota che la transconduttanza g_m in debole inversione risulta indipendente dalla lunghezza di canale, mentre dall'equazione (2.12) g_{ds} risulta inversamente proporzionale ad essa. Conseguentemente, il guadagno intrinseco aumenta all'aumentare di L_G . In forte inversione la transconduttanza di canale g_m è data dalla (2.3), mentre la conduttanza di uscita è espressa da:

$$g_{ds} = \lambda I_D, \quad (2.13)$$

dove λ può essere approssimata come:

$$\lambda \approx \frac{1}{V_{DS} - V_{DS_{SAT}}} \cdot \frac{\Delta L}{L_G}, \quad (2.14)$$

con ΔL che rappresenta la riduzione della lunghezza di canale e vale:

$$\Delta L \approx \sqrt{\frac{2\epsilon_s}{qN_A} (V_{DS} - V_{DS_{SAT}})}. \quad (2.15)$$

Il guadagno intrinseco di tensione risulta dunque, in questa regione operativa, proporzionale a $I_D^{-1/2}$. La figura 2.20 mostra, tuttavia, una pendenza della

curva più ripida in regione di forte inversione a causa della saturazione della velocità dei portatori e ad altri effetti di canale corto. In questo caso, infatti, come si evidenzia dalla (2.6), la transconduttanza g_m è indipendente dalla corrente di drain.

L'impatto dello scaling sul guadagno intrinseco A_{Vi} è valutato nella figura 2.21, che mostra la dipendenza di questo parametro dalla lunghezza di gate per dispositivi NMOS appartenenti a 4 differenti nodi tecnologici, inclusi i transistori sotto misura. I dispositivi cui si riferiscono i dati in figura sono polarizzati in maniera tale da lavorare tutti col medesimo coefficiente di in-

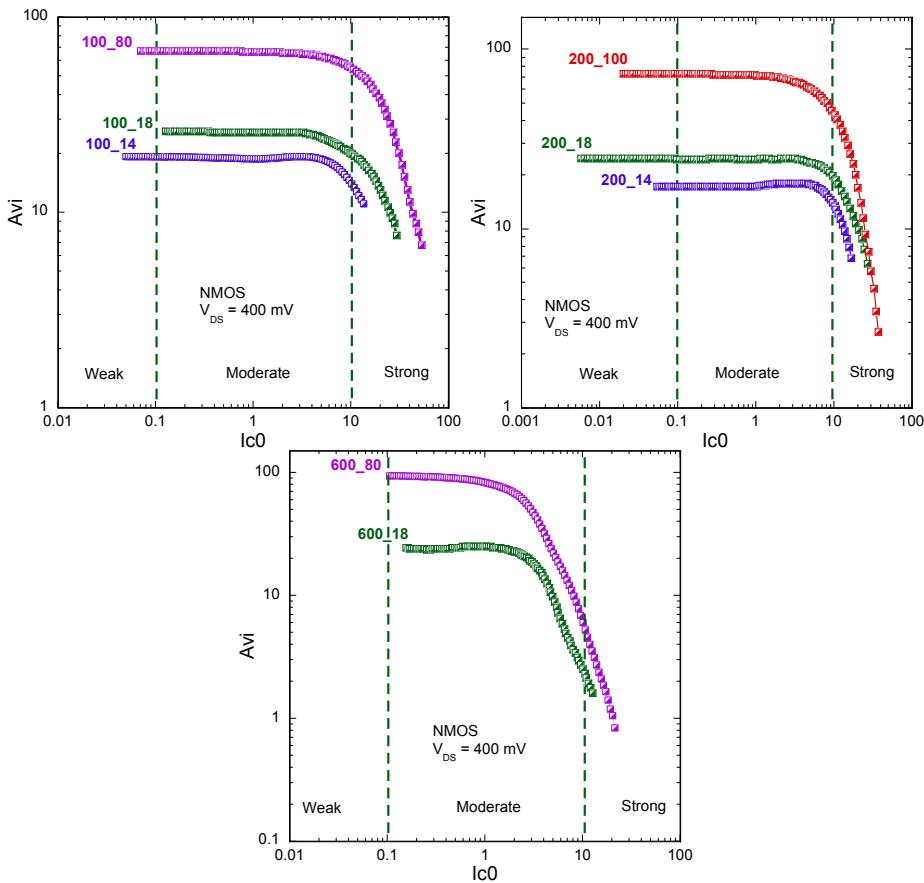


Figura 2.20: guadagno intrinseco A_{Vi} in funzione della lunghezza di gate L_G per dispositivi NMOS appartenenti a diversi nodi tecnologici e polarizzati a $I_{C0} = 0.1$.

versione, $I_{C0} = 0.1$, vale a dire al confine tra debole e moderata inversione, dove il guadagno è vicino al suo valore asintotico. Infine è da notare che per lunghezze di canale minime il guadagno intrinseco risulta uguale per tutti i nodi tecnologici. Mantenere A_{Vi} costante con lo scaling è considerato una delle principali sfide nel progetto delle tecnologie nanometriche [29]. È possibile infatti dimostrare che [24]:

$$\frac{g_m}{g_{ds}} = g_m \cdot r_0 \propto \alpha \cdot L_G, \quad (2.16)$$

dove α è il fattore di scaling. Di conseguenza se L_G decresce di α il guadagno intrinseco di tensione rimane costante.

Per quanto riguarda i dispositivi a canale P non è stato possibile effettuare questo tipo di caratterizzazione a causa della presenza, nella loro corrente di drain, di contributi di corrente spuri, probabilmente provenienti dalle strutture di protezione da scarica eletrostatica e/o dagli altri dispositivi presenti nella struttura di test (che come già detto, hanno terminali di source e drain in comune). Tali contributi alterano il comportamento dei transistori a canale P nella regione di debole inversione, rendendo vano il tentativo di estrarre informazioni circa il loro guadagno intrinseco.

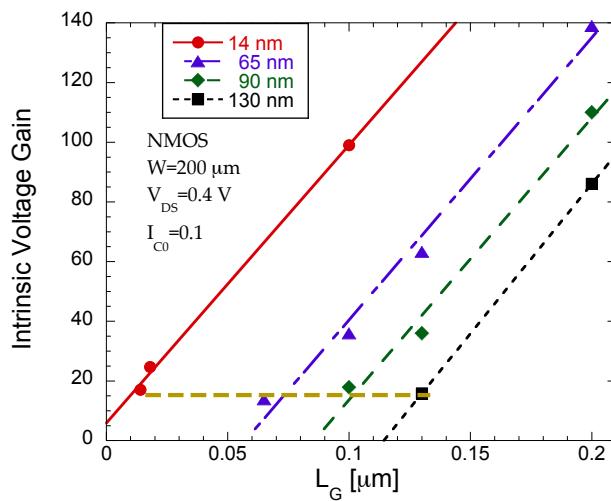


Figura 2.21: guadagno intrinseco A_{Vi} in funzione della lunghezza di gate L_G per dispositivi NMOS appartenenti a diversi nodi tecnologici e polarizzati nella regione di inversione $I_{C0} = 0.1$.

2.2.3 Corrente di perdita di gate

In un processo CMOS che utilizza lunghezze di canale minime dell'ordine della decina dei nanometri non è possibile trascurare la corrente di gate, che è fondamentalmente determinata dal passaggio di portatori di carica da gate a source, drain e canale e vice versa, in modo casuale ma continuo e dipendente dalla tensione di polarizzazione, attraverso il sottilissimo (dell'ordine di qualche nanometro) strato di isolante che separa il gate dal canale. Questo fenomeno è favorito da effetti di *tunneling* assistito da trappole o, nel caso di strati di isolante molto sottili, diretto [30]. Questo comporta un aumento del consumo di potenza statica, particolarmente critico nel caso si utilizzi il dispositivo in applicazioni digitali, o un degrado delle prestazioni di rumore a causa del contributo aggiuntivo di tipo 1/f e granulare nella corrente di gate, che può diventare particolarmente significativo in alcune applicazioni analogiche. La densità di corrente di gate è una figura di merito importante e comunemente utilizzata per valutare la riduzione dello spessore dell'ossido e il conseguente aumento della corrente I_G . Si tratta di una corrente normalizzata all'area di gate del dispositivo e misurata cortocircuitando drain, source e bulk. Nei dispositivi studiati tuttavia non è stato possibile effettuare questo tipo di caratterizzazione. La configurazione dei dispositivi nelle strutture di test, con source e gate in comune, non ha consentito infatti di distinguere i contributi di corrente di gate dei singoli transistori.

2.2.4 Estrazione della tensione di soglia

La tensione di soglia V_{TH} di un transistore MOS è definita come quella tensione tra gate e bulk per la quale la popolazione di minoritari all'interfaccia è uguale alla popolazione di maggioritari nel bulk. Questa definizione non può essere utilizzata direttamente per l'estrazione di V_{TH} , per la quale in realtà ci si affida tipicamente alla elaborazione della caratteristica corrente tensione dei dispositivi. Il valore di V_{TH} deve essere determinato in maniera precisa ed affidabile e la sua estrapolazione risulta essere sempre più complessa a causa della continua riduzione dello spessore dell'ossido e della tensione di alimentazione. In letteratura sono proposte differenti tecniche di estrazione della tensione di soglia, tra le quali, una delle più utilizzate è quella della conduttanza massima. Questo metodo, denominato *Transconductance Change Method, TCM* [31], definisce la tensione di soglia come la tensione di gate V_{GS} corrispondente al picco massimo della derivata della transconduttanza g_m rispetto alla tensione di gate ed è valido per bassi valori della tensione V_{DS} . Esso risulta adeguato per i dispositivi MOSFET caratterizzati da un sottile dielettrico di gate o da

un *Ultra Thin Body* SOI o per i transistor a gate multiplo. In dispositivi con queste caratteristiche fisiche e geometriche, le tecniche di estrazione lineare, mediante le quali il valore di V_{TH} è ottenuto dall'extrapolazione lineare della curva I_D-V_{GS} [32] o del rapporto $I_D/g_m^{1/2}$ [33] in funzione della tensione V_{GS} , potrebbero portare a valori sovrastimati della tensione di soglia [31]. I limiti di questi metodi sono dovuti alla loro dipendenza dagli effetti di degradazione della mobilità dei portatori e dalla resistenza serie, nonché dall'assunzione che in conduzione la carica mobile nello strato di inversione, Q_{inv} , dipenda linearmente dalla tensione di gate. In dispositivi nanometrici caratterizzati da un dielettrico di gate estremamente sottile e paragonabile allo spessore dello strato di inversione, la carica Q_{inv} ha una dipendenza approssimativamente lineare dalla tensione di gate solo per campi elettrici perpendicolari elevati [34]. Si rende quindi necessaria l'adozione di tecniche più sofisticate per la corretta estrazione del valore della tensione di soglia.

Un metodo alternativo al TCM definisce la tensione di soglia come la tensione di gate alla quale la derivata seconda del logaritmo della corrente di drain rispetto alla tensione V_{GS} ($d^2(\log I_{DS})/dV_{GS}^2$) raggiunge il suo valore minimo [35]. Nel caso di un dispositivo a canale N, in debole inversione la corrente di drain è approssimata mediante una relazione esponenziale che può essere scritta come:

$$I_{DS} = \mu_n C_{ox} \frac{W}{L} \frac{1}{m} (n_{sub} V_t)^2 \exp \left(\frac{V_G - V_{TH} - n_{sub} V_t}{n_{sub} V_t} \right) \cdot \left[1 - \exp \left(-\frac{m V_{DS}}{n_{sub} V_t} \right) \right], \quad (2.17)$$

con m circa uguale a 1 nei transistori appartenenti alle tecnologie più recenti. La ragione fisica per cui l'espressione della corrente di sottosoglia è completamente diversa dalla relazione che caratterizza il funzionamento del MOS in conduzione è dovuta ai differenti meccanismi di trasporto che predominano in queste due regioni di lavoro. In debole inversione la carica mobile nel canale è trascurabile rispetto alla carica fissa. Essa non contribuisce pertanto alla densità di carica spaziale ed il potenziale superficiale non varia lungo il canale se non nelle immediate vicinanze del drain. Per questa ragione il meccanismo prevalente di trasporto è la diffusione e l'intensità della corrente è limitata dalla barriera di potenziale che si genera all'ingresso del canale, analogamente a quanto avviene in un transistore bipolare². Da ciò discende la natura esponenziale della caratteristica. In forte inversione la corrente è invece dovuta alla sola componente di drift dei portatori mobili e viene descritta mediante un'espressione polinomiale che dipende dalla regione operativa. In regione lineare

²Ipotizzando che il transistore sia un NMOS, in assenza del canale conduttivo le regioni di source (n^+), di substrato (p) e di drain (n^+) formano un transistore bipolare parassita.

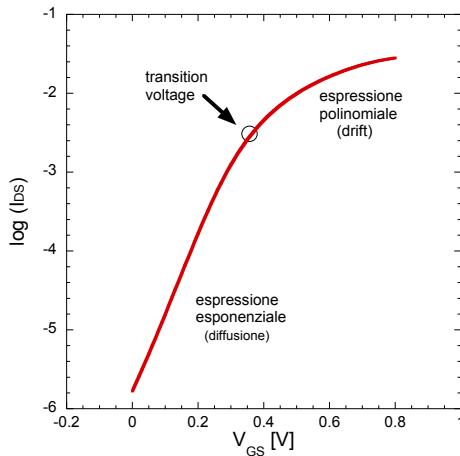


Figura 2.22: $\log(I_D)$ in funzione di V_{GS} .

la corrente di drain I_{DS} è espressa da:

$$I_{DS} = \mu_n C_{ox} \frac{W}{L} \left[(V_{GS} - V_{TH}) V_{DS} - \frac{V_{DS}}{2} \right]; \quad (2.18)$$

in saturazione da:

$$I_{DS} = \frac{1}{2} \mu_n C_{ox} \frac{W}{L} (V_{GS} - V_{TH})^2. \quad (2.19)$$

La tensione di soglia può dunque essere ottenuta ricavando la tensione di gate corrispondente al punto di transizione che delimita l'espressione polinomiale da quella in forma esponenziale, come rappresentato in figura 2.22. Si nota che nel punto di transizione tra la debole e la forte inversione la curva rappresentata cambia pendenza ed è naturale pensare che la derivata di tale curva diminuisca bruscamente in corrispondenza del punto cercato. È possibile affermare che la tensione di soglia è la tensione di gate che corrisponde alla massima variazione possibile della pendenza della curva $\log(I_{DS})-V_{GS}$ e che coincide dunque con il valore minimo della derivata seconda del logaritmo della corrente di drain rispetto a V_{GS} . La tensione di soglia V_{TH} è così definita come la tensione di gate alla quale la componente di diffusione risulta uguale alla componente di drift.

I punti seguenti schematizzano i passi operativi per l'estrapolazione della V_{TH} secondo il metodo del minimo della derivata seconda del logaritmo (SDLM, *second difference of the logarithm of the drain current minimum*):

- si calcola la derivata prima del logaritmo della corrente di drain rispetto alla tensione di gate, $d(\log I_{DS})/dV_{GS}$ (figura 2.23(a));

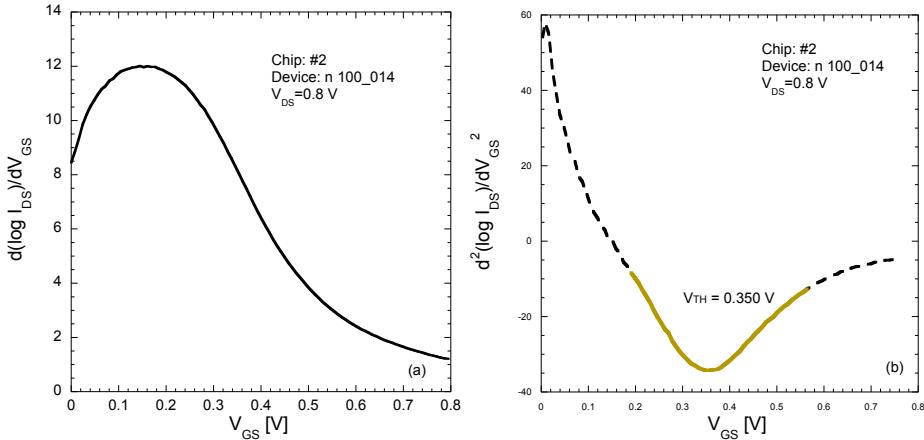


Figura 2.23: rappresentazione grafica del metodo SDLM per un FinFET con $W/L=100 \mu m/14 nm$.

- si calcola la derivata seconda della corrente di drain rispetto alla tensione di gate, $d^2(\log I_{DS})/dV_{GS}^2$ (figura 2.23(b));
- la tensione di soglia è la tensione di gate corrispondente al valore minimo di $d^2(\log I_{DS})/dV_{GS}^2$.

Il metodo descritto risulta formalmente equivalente al metodo che si basa sul picco massimo della derivata del rapporto g_m/I_{DS} rispetto alla tensione gate-sorce [34]. Infatti la quantità $d(\log I_{DS})/dV_{GS}$ corrisponde al rapporto g_m/I_{DS} :

$$\frac{g_m}{I_{DS}} = \frac{1}{I_{DS}} \frac{dI_{DS}}{dV_{GS}} = \frac{d(\log I_{DS})}{dV_{GS}}. \quad (2.20)$$

Calcolando la derivata della (2.20) si ottiene:

$$\frac{d(g_m/I_{DS})}{dV_{GS}} = \frac{d^2(\log I_{DS})}{dV_{GS}^2}. \quad (2.21)$$

In definitiva la tensione di soglia corrisponde al valore massimo di $-\frac{d(g_m/I_{DS})}{dV_{GS}}$. Per l'estrazione del valore della tensione di soglia, in questo tesi, si utilizza il metodo della derivata seconda del logaritmo. I risultati ottenuti vengono comunque confrontati con il metodo della transconduttanza massima applicato in figura 2.24 ad un dispositivo FinFET con $W/L = 100 \mu m/14 nm$. Si nota come il valore di picco di $d g_m/dV_{GS}$ all'aumentare di V_{DS} , corrisponda

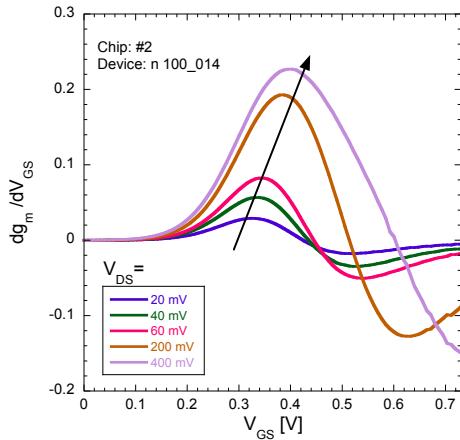


Figura 2.24: dg_m/dV_{GS} in funzione della tensione di gate V_{GS} per differenti valori della tensione di drain V_{DS} .

a valori sempre più elevati della tensione gate-source. Ciò implica un errore nell'estrazione della tensione di soglia dipendente dalla tensione di drain applicata. In effetti, in accordo con [36], il metodo del massimo di dg_m/dV_{GS} deve essere applicato a basse tensioni di drain mentre quello della derivata seconda ($d(\log I_{DS})/dV_{GS}$) è impiegato in regione di saturazione. Con questo accorgimento le due tecniche di estrazione forniscono un valore della tensione di soglia simile, come mostrato in figura 2.25. La figura 2.26 mostra la distribuzione dei valori di soglia estratti per i diversi dispositivi. L'analisi dei dati ha fornito un valore di 330 mV per la tensione di soglia dei dispositivi a canale N e di 280 mV per i dispositivi a canale P.

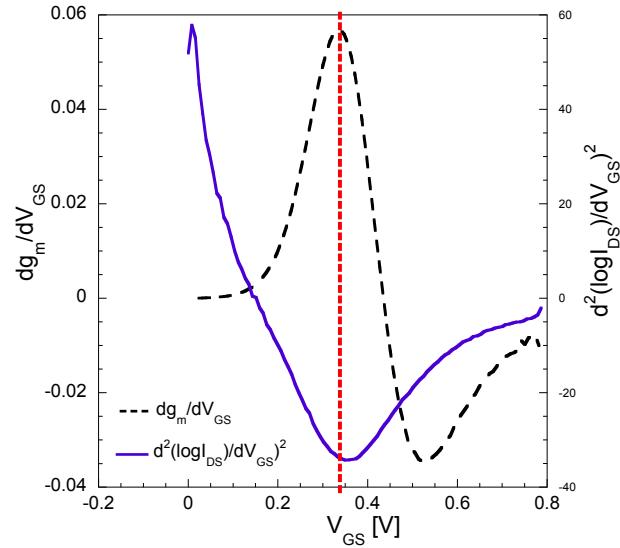


Figura 2.25: confronto tra due diversi metodi di estrazione del valore della tensione di soglia.

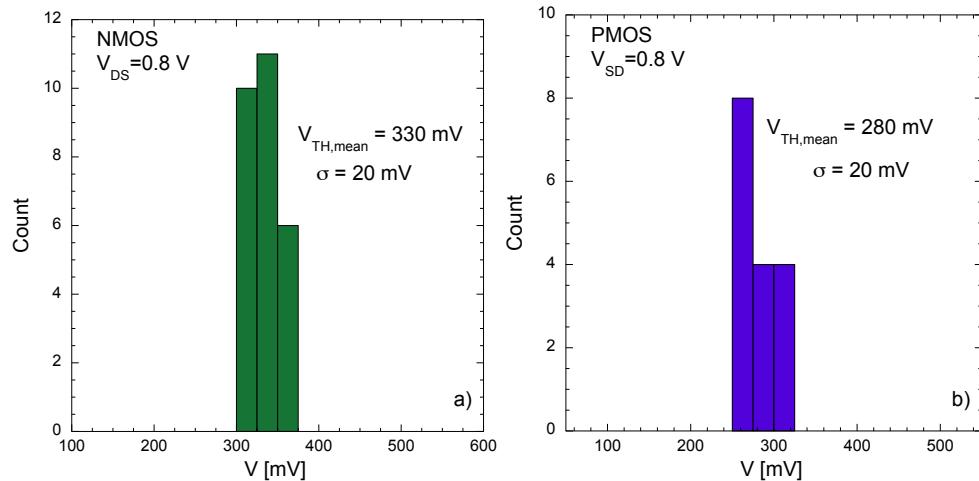


Figura 2.26: distribuzione dei valori estratti per la tensione di soglia per i dispositivi a) NMOS e b) PMOS.

2.2.5 Pendenza della caratteristica $I_D - V_{GS}$ in sottosoglia

Dall'osservazione delle curve $I_D - V_{GS}$ mostrate precedentemente è possibile notare che la corrente non crolla bruscamente a zero per $V_{GS} < V_{TH}$. Ciò equivale a dire che il transistor è già parzialmente in conduzione per tensioni minori della tensione di soglia, dove l'andamento della corrente è fornito dalla (2.17). La pendenza con cui la caratteristica $\log(I_D) - V_{GS}$ attraversa la regione di sottosoglia fornisce un'indicazione sulle proprietà del dispositivo in condizioni di spegnimento: quanto maggiore è la pendenza, tanto più rapidamente il dispositivo si sposta, al diminuire di V_{GS} , da una condizione di debole conduzione ad una di conduzione minima, con correnti di drain trascurabili. Questo aspetto è di notevole interesse in applicazioni digitali, in cui è sicuramente auspicabile che un dispositivo nominalmente spento ma caratterizzato comunque, anche in quello stato, da una conduzione di corrente non nulla, dissipi la minima quantità di potenza possibile. Il problema è acuito dall'evoluzione tecnologica, in cui, alla riduzione delle dimensioni e della tensione di alimentazione si tenderebbe a far corrispondere una diminuzione della tensione di soglia. Per definizione, la pendenza di sottosoglia S fornisce la variazione della tensione di gate che dà luogo ad una variazione di un fattore 10 della corrente di sottosoglia e si esprime in mV/dec. Nel capitolo 1 si è visto come le nuove tecnologie nanometriche (ed in particolare FD-SOI e FinFET) siano caratterizzate da una pendenza di sottosoglia minore rispetto ai processi bulk CMOS convenzionali. In base alla definizione fornita sopra:

$$S = [d \log(I_{DS}) / dV_{GS}]^{-1} = \left(\frac{n_{sub}k_B T}{q} \right) \ln 10. \quad (2.22)$$

In un transistor ideale $n_{sub} = 1$ e $S = (K_B T/q) \ln 10 \simeq 60$ mv/dec a temperatura ambiente, il che vuol dire che la corrente di sottosoglia diminuisce di un fattore 10 per una riduzione della tensione gate-source pari a 60 mV. Il valore della pendenza di sottosoglia può essere estratto a partire dalla curva $I_D - V_{GS}$ in scala semilogaritmica mediante un fit lineare, come mostrato in figura 2.27 per un FinFET a canale N e per uno a canale P con $W/L = 100\mu m/18nm$. Nella figura 2.28 viene rappresentato il valore di S estrapolato per le diverse lunghezze di canale L_G , per dispositivi NMOS e PMOS. Dai grafici si nota che la pendenza S aumenta con il diminuire di L_G , aumento determinato dagli effetti di canale corto. Il basso valore di S in generale nei dispositivi DG-MOSFETs è dovuto alla presenza di una bassa concentrazione di drogante nelle regioni indicate in figura 2.29 come L_{eS} e L_{eD} . In debole inversione, vicino al gate, la densità di elettroni n è bassa e dipende direttamente dalla tensione applicata al gate. Quindi il controllo del gate sugli elettroni si

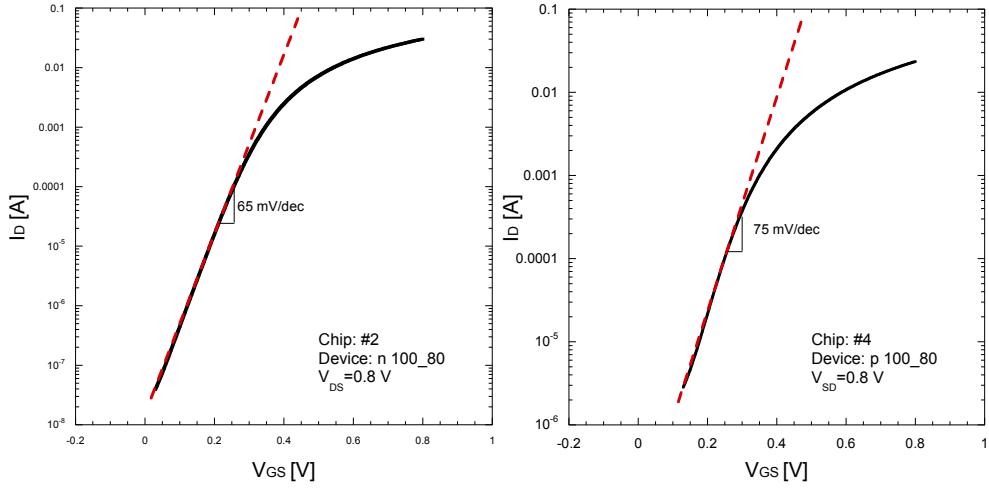


Figura 2.27: estrazione della pendenza di sottosoglia S .

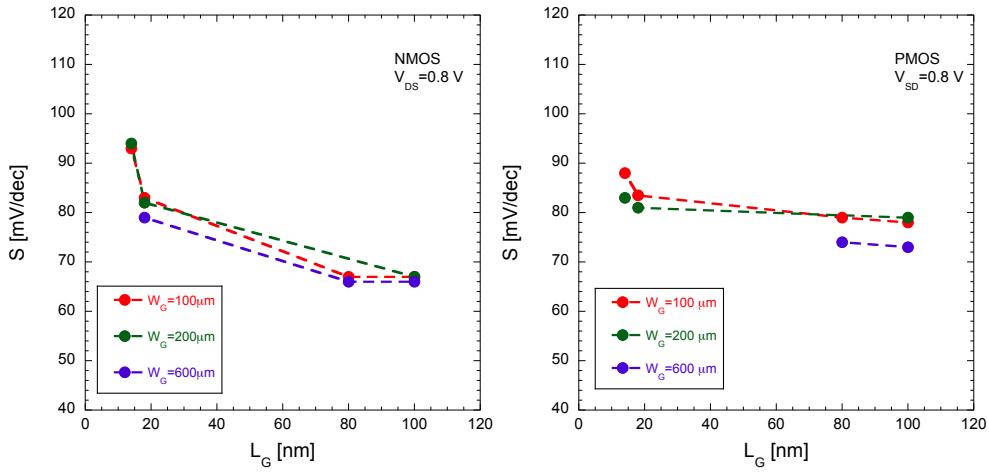


Figura 2.28: dipendenza di S dalla la lunghezza di canale L_G .

estende anche nelle regioni adiacenti. Ne deriva che la lunghezza di canale effettiva L_{eff} risulta maggiore della lunghezza di gate nominale L_G e può essere espressa come:

$$L_{eff(weak)} \approx L_G + L_{eS} + L_{eD}, \quad (2.23)$$

dove L_{eS} e L_{eD} dipendono dalla lunghezza di Debye ($L_D \propto 1/\sqrt{n}$) e dunque dalla densità di elettroni. In forte inversione, invece, n è alto e conseguentemente

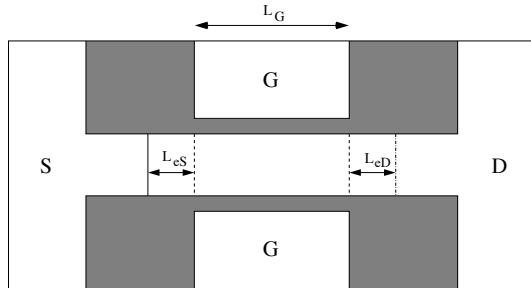


Figura 2.29: struttura schematizzata di un DG-MOSFET che indica le porzioni non drogata L_{eS} e L_{eD} .

mente L_D è più corta ed il controllo del canale da parte del terminale di gate risulta limitato alla lunghezza:

$$L_{eff} \approx L_G. \quad (2.24)$$

A fini comparativi in figura 2.30 la pendenza di sottosoglia estratta dalla tecnologia studiata viene messa a confronto con quella di dispositivi bulk CMOS appartenenti ad un nodo tecnologico precedente. Si nota che a parità di lun-

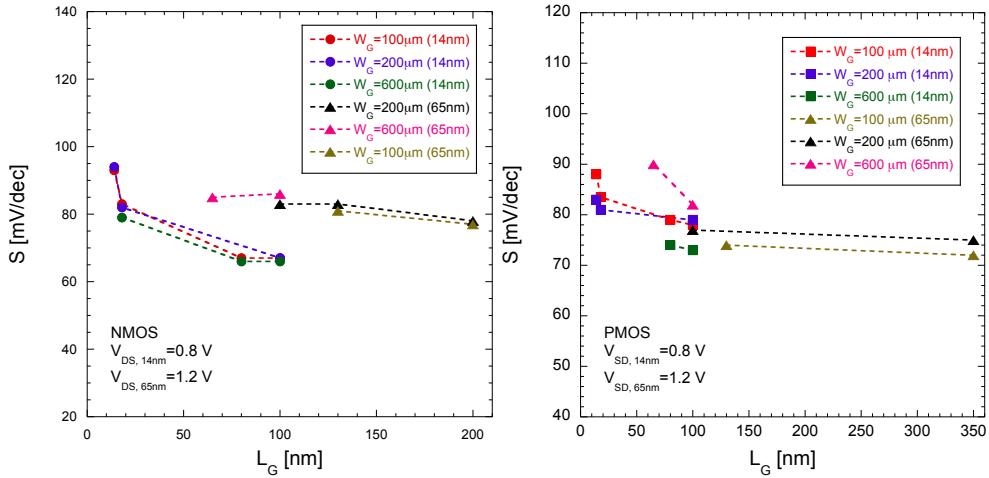


Figura 2.30: confronto della pendenza di sottosoglia S tra i dispositivi realizzati in tecnologia da 14 nm e quelli appartenenti ad una tecnologia da 65 nm.

ghezza di gate, la pendenza di sottosoglia dei dispositivi FinFET realizzati in tecnologia da 14 nm è inferiore, nel caso di dispositivi a canale N, a quella dei dispositivi realizzati in tecnologia CMOS da 65 nm. Nel caso dei dispositivi a canale P, il dispositivo con $W = 200 \mu m$ della tecnologia a 65 nm ha un valore di S inferiore rispetto al transistore FinFET con W corrispondente.

Capitolo 3

Caratterizzazione di rumore in dispositivi FinFET

In questo capitolo vengono dapprima fornite le nozioni fondamentali riguardanti i modelli e i parametri di rumore per dispositivi CMOS nanometrici. Successivamente si analizza la strumentazione utilizzata per le misure di rumore in dispositivi FinFET, in termini generali e nel caso specifico del setup di misura utilizzato in questo lavoro. Il capitolo prosegue e si conclude con l'analisi dei dati sperimentali.

3.1 Sorgenti equivalenti di rumore nei circuiti lineari

Il rumore elettronico per definizione è un processo stocastico continuo a parametro continuo, il tempo, che ha origine da fenomeni fisici inerenti al funzionamento dei componenti elettronici. Una sorgente di rumore può essere rappresentata mediante un generatore di tensione caratterizzato da una densità spettrale di potenza, misurata in V^2/Hz (si parla, in questo caso, di rappresentazione alla Thevenin), oppure mediante un generatore di corrente descritto da una densità spettrale di rumore, misurata in A^2/Hz (rappresentazione alla Norton). Considerando la rete lineare con un generatore di rumore al proprio ingresso mostrata in figura 3.1, è possibile calcolare la densità spettrale di rumore all'uscita della rete mediante l'espressione:

$$S_u(\omega) = S_i(\omega) \cdot |T(j\omega)|^2 \quad (3.1)$$

dove $S_i(\omega)$ e $S_u(\omega)$ rappresentano, rispettivamente, la densità spettrale di rumore all'ingresso e all'uscita della rete lineare considerata, mentre $T(s)$ è la

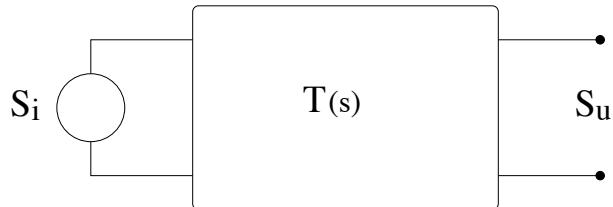


Figura 3.1: rappresentazione schematica di una rete lineare con sorgente di rumore al proprio ingresso.

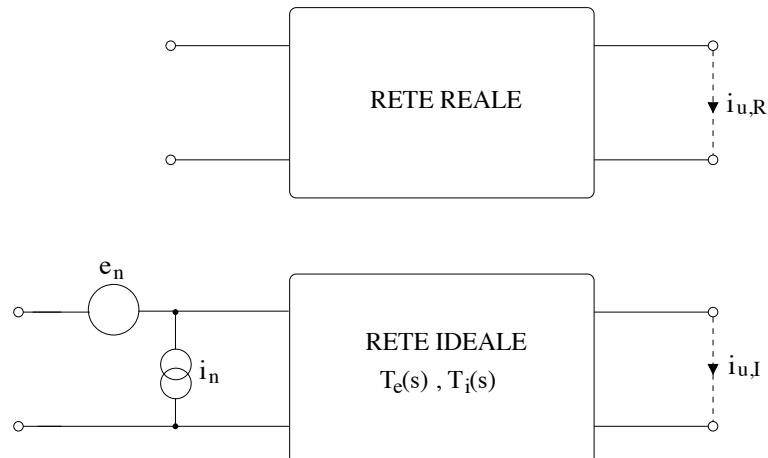


Figura 3.2: rappresentazione del rumore in un circuito elettronico mediante generatori equivalenti riferiti all'ingresso.

sua funzione di trasferimento.

Un circuito lineare reale, quindi rumoroso, può essere sempre rappresentato mediante un circuito ideale privo di rumore con un generatore di tensione, o serie, ed un generatore di corrente, o parallelo, connessi all'ingresso del circuito, come mostrato in figura 3.2. Tali generatori vengono detti generatori equivalenti in quanto producono all'uscita del sistema ideale lo stesso effetto prodotto dalle sorgenti di rumore interne al circuito reale. Indicando con $T_e(s)$ la funzione di trasferimento tra il generatore equivalente serie e_n e l'uscita e con $T_i(s)$ quella tra il generatore equivalente parallelo i_n e l'uscita, si determina la densità spettrale di rumore associata a questi generatori come segue:

1. nel caso del generatore equivalente di tensione
 - si cortocircuitano gli ingressi sia nel circuito reale, sia nel circuito ideale;
 - si eguaglano le densità spettrali di rumore alle due uscite;
2. nel caso del generatore equivalente di corrente
 - si aprono gli ingressi sia nel circuito reale sia nel circuito ideale;
 - si eguaglano le densità spettrali di rumore alle due uscite.

Prendendo come grandezza di uscita la corrente di cortocircuito, i_u , si ha rispettivamente:

$$\frac{d\overline{i_{u,R}^2}}{df} = \frac{d\overline{i_{u,I}^2}}{df} = \frac{d\overline{e_n^2}}{df} |T_e(j\omega)|^2 \Rightarrow \frac{d\overline{i_n^2}}{df} = \frac{d\overline{i_{u,R}^2}}{df} \frac{1}{|T_e(j\omega)|^2}; \quad (3.2)$$

$$\frac{d\overline{i_{u,R}^2}}{df} = \frac{d\overline{i_{u,I}^2}}{df} = \frac{d\overline{i_n^2}}{df} |T_i(j\omega)|^2 \Rightarrow \frac{d\overline{i_n^2}}{df} = \frac{d\overline{i_{u,R}^2}}{df} \frac{1}{|T_i(j\omega)|^2}. \quad (3.3)$$

3.1.1 Parametri e modelli di rumore per dispositivi CMOS nanometrici

I transistor ad effetto di campo sono caratterizzati dalla presenza di rumore elettronico di varia natura, che, come visto precedentemente, può essere modellizzato mediante due sorgenti di rumore, serie e parallelo, riferite al gate. Se la corrente di gate è trascurabile sarà in generale pure trascurabile il contributo di rumore parallelo. Il generatore di rumore serie avrà invece densità spettrale di rumore:

$$\frac{d\overline{e_n^2}}{df} = S_e(f) = S_w + S_{\frac{1}{f}}(f). \quad (3.4)$$

Nella trattazione che segue si analizzano le componenti di rumore in un MOSFET con particolare riferimento ai contributi di rumore nella corrente di canale, assumendo trascurabili i contributi di rumore nella corrente di gate [37].

Rumore bianco

Il primo addendo della (3.4), S_w , rappresenta i contribuiti di rumore indipendenti dalla frequenza, in particolar modo, il rumore termico di canale ed il rumore dovuto alle resistenze parassite di gate e di substrato. Generalmente

il rumore termico di canale costituisce il contributo più significativo.

In un dispositivo MOSFET, per esempio a canale N, l'applicazione di una differenza di potenziale tra gate e substrato superiore alla tensione di soglia, determina la formazione al di sotto dell'ossido di gate di uno strato di inversione dove la concentrazione di elettroni raggiunge un valore superiore alla concentrazione di lacune nella zona neutra. Tale regione, che costituisce il canale del dispositivo, viene modulata dalla tensione gate-source e si comporta pertanto come una resistenza variabile nella quale i portatori sono soggetti ad agitazione termica. L'interpretazione di Nyquist, circa il rumore elettronico in un resistore, implica che il canale del MOSFET è sede di rumore termico, detto appunto rumore termico di canale.

La densità spettrale di potenza associata a questo contributo di rumore, valida in qualsiasi regione di funzionamento del transistor e modellizzata mediante un generatore di corrente parallelo, come mostrato in figura 3.3(a), può essere espressa dalla seguente densità spettrale di potenza:

$$S_{Id} = \frac{4K_B T}{L^2 I_D} \int_0^{V_{DS}} g^2(V') dV' \quad (3.5)$$

dove $g(V')$ rappresenta la conduttanza di canale misurata nel punto in cui la tensione rispetto al source vale V' .

L'espressione (3.5) può essere semplificata nel modo seguente:

$$S_{Id} = 4K_B T \gamma g_{d0} \quad (3.6)$$

con

$$\gamma = \frac{1}{g_{d0} L^2 I_D} \int_0^{V_{DS}} g^2(V') dV' \quad (3.7)$$

dove g_{d0} è la conduttanza di canale misurata mantenendo la tensione drain-source nulla. Infine, la (3.6), se riferita al gate del transistor, come mostrato in figura 3.3(b) assume la seguente espressione:

$$S_w = 4K_B T \frac{\alpha}{g_m} \quad (3.8)$$

con $\alpha = \gamma g_{d0} / g_m$.

Tipicamente, la densità spettrale di rumore termico nella corrente di drain riferita all'ingresso del dispositivo viene espressa come:

$$S_w = \frac{4K_B T \Gamma}{g_m} \quad (3.9)$$

dove Γ è il cosiddetto coefficiente di rumore termico di canale, il cui valore dipende dalle condizioni di polarizzazione del dispositivo e varia tra 1/2 in

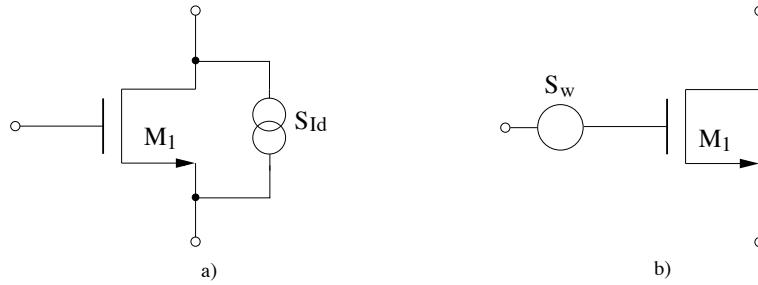


Figura 3.3: rumore termico di canale di un MOSFET modellizzato come a) generatore di corrente parallelo e b) generatore di tensione serie.

debole inversione e $2/3$ in forte inversione. Dalle (3.8) e (3.9) si deduce che $\alpha = \Gamma$. Inoltre, la transconduttanza di canale, nel caso di dispositivi a canale lungo in saturazione, coincide con g_{d0} . Questo si ricava applicando la definizione di g_m e di g_{d0} , tenendo conto delle (2.19) e (2.18), si ricava semplicemente che $g_m = g_{d0}$.

Le equazioni viste tuttavia non risultano corrette nel caso in cui il dispositivo abbia una lunghezza di canale dell'ordine della frazione di micron, poiché in questa situazione si manifestano gli effetti di canale corto già trattati nel capitolo 1. Questi fenomeni comportano un aumento del rumore termico di canale, in quanto [23]:

- il rapporto g_{d0}/g_m è diverso da uno a causa della dipendenza di g_{d0} dal campo elettrico trasversale e dalla dipendenza di g_m dalla modulazione della lunghezza di canale;
- l'espressione di γ deve essere modificata a causa dell'effetto dovuto ai portatori caldi, alla modulazione del canale ed alla dipendenza della mobilità dei portatori dal campo elettrico longitudinale¹ e trasversale.

Considerando questi effetti, per i dispositivi nanometrici, S_w si discosta dalla (3.9), e viene espressa dalla seguente equazione [23]:

$$S_w = 4K_B T \frac{\alpha_w n_{sub} \gamma}{g_m} \quad (3.10)$$

¹Al diminuire della minima lunghezza di canale il campo elettrico longitudinale aumenta ($E = V_{DS}/L$). Valori elevati di E comportano una diminuzione della mobilità e la velocità dei portatori tende a raggiungere un valore di saturazione $v_{sat} = \mu_0 E_C$ (con μ_0 mobilità a bassi campi ed E_C campo critico).

dove $\alpha_w > 1$ è il fattore di rumore in eccesso che tiene conto degli effetti di canale corto, che può essere estrapolato da misure sperimentali, ed n_{sub} è un coefficiente proporzionale al reciproco della pendenza della caratteristica $I_D - V_{GS}$ nella regione di sottosoglia, già trattato nel capitolo precedente. Esistono in letteratura diversi modelli che sono stati proposti per descrivere il coefficiente di rumore termico di canale γ , che può essere espresso mediante la seguente relazione [38]:

$$\gamma = \frac{1}{1 + \frac{I_D L}{I_Z^* W}} \left[\frac{1}{2} + \frac{2 I_D L}{3 I_Z^* W} \right]. \quad (3.11)$$

La corrente di drain caratteristica normalizzata, I_Z^* , come visto nel capitolo 2, può essere estratta mediante misure statiche.

Il rumore termico di canale può essere valutato in termini di una resistenza equivalente, R_{eq} , data da:

$$R_{eq} = \frac{S_w}{4 K_B T} = \alpha_w \frac{n_{sub} \gamma}{g_m}. \quad (3.12)$$

Ricorrendo al coefficiente di inversione I_{C0} , definito nell'espressione (2.8) e considerando l'equazione (3.11), la resistenza equivalente di rumore termico si può riscrivere come:

$$R_{eq} = R_{eq,0} \frac{L}{W} f_{R_{eq}}(I_{C0}) \quad (3.13)$$

dove:

$$R_{eq,0} = \alpha_w \frac{n_{sub}^2 V_t}{I_Z^*} \quad (3.14)$$

risulta dipendente dai parametri tecnologici (α_w, n_{sub}, I_Z^*). Il termine $f_{R_{eq}}(I_{C0})$, nella (3.13) è, invece, una funzione che esprime la dipendenza del rumore termico dal livello di inversione e non dipende dalla tecnologia [38].

Rumore flicker

Il secondo termine dell'equazione (3.4) rappresenta il rumore $1/f$ nella corrente di canale, e viene usualmente descritto da differenti modelli che tengono conto o delle fluttuazioni della densità dei portatori (modello di McWhorter) o dalle variazioni nella mobilità di elettroni e lacune (modello di Hooge). Nel modello proposto da McWhorter [39] la variazione della corrente di canale è causata da trappole localizzate all'interfaccia tra ossido e canale dovute a contaminazioni e difetti del cristallo, come mostrato nella figura 3.4. I portatori che attraversano la regione di canale, infatti, sono caratterizzati da una probabilità non nulla

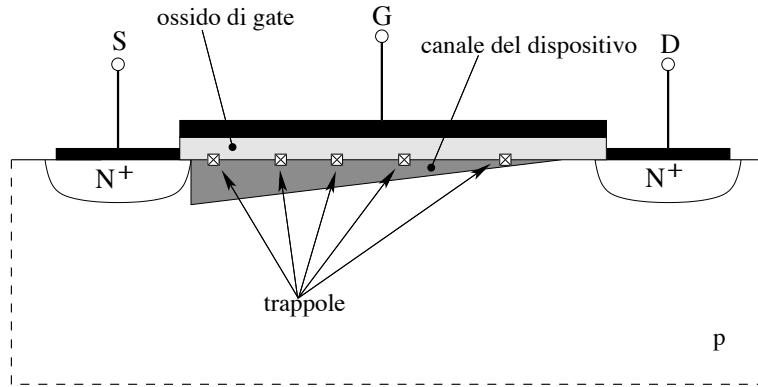


Figura 3.4: sezione trasversale di un dispositivo MOSFET a canale N.

di essere catturati o rilasciati da queste trappole, fatto che determina una fluttuazione della corrente di canale. Le costanti di tempo associate a questo processo danno luogo ad un segnale di rumore con energia concentrata a basse frequenze. Pertanto, la densità spettrale di rumore risulta essere inversamente proporzionale alla frequenza. Il modello di McWorther prevede una sorgente di tensione di rumore equivalente al gate del dispositivo caratterizzata da una densità spettrale di potenza che può essere descritta dalla seguente relazione:

$$S_{\frac{1}{f}}(f) = \frac{K_f}{C_{ox}^2 WL} \frac{1}{f^{\alpha_f}} \quad (3.15)$$

dove K_f è un parametro dipendente dalla tecnologia e α_f tiene conto della dipendenza dalla frequenza (in scala logaritmica il parametro α_f è legato alla pendenza della densità spettrale di potenza associata al rumore *flicker*, ed è idealmente uguale ad 1).

Il modello basato invece sulla teoria di Hooge [40], prevede che le variazioni nella mobilità dei portatori siano causate dallo scattering dei fononi nel bulk del semiconduttore al di sotto dell'interfaccia $Si-SiO_2$. Tali fluttuazioni danno origine ad una variazione della corrente di drain. In accordo con questo modello il contributo di rumore flicker mostra un aumento con la tensione di overdrive e può essere descritto dalla relazione:

$$S_{\frac{1}{f}}(f) = \frac{K_f(V_{OV})}{C_{ox}WL} \frac{1}{f^{\alpha_f}} \quad (3.16)$$

dove $V_{OV} = V_{GS} - V_{TH}$ è la tensione di overdrive e $K_f(V_{OV})$ è un parametro di processo. Poiché la densità spettrale di potenza per il rumore *flicker* osservato

nei dispositivi MOS a canale N, risulta spesso costante con il livello di inversione, il modello di McWhorter è, in genere, associato ai dispositivi NMOS; al contrario, poiché per i dispositivi a canale P la densità spettrale di rumore aumenta con il livello di inversione, ad essi viene associato il modello di Hooke. Quindi il contributo di rumore $1/f$ può essere approssimato utilizzando le espressioni (3.15) e (3.16) rispettivamente per gli NMOS e per i PMOS.

Nei dispositivi a canale P, solitamente la $S_{\frac{1}{f}}$ è minore rispetto a quella dei dispositivi NMOS appartenenti alla medesima tecnologia, a pari dimensione e condizioni di polarizzazione. Il processo di *scaling* dei dispositivi, studiato nel capitolo 1, dovrebbe portare apparentemente ad una diminuzione del rumore flicker, dovuta principalmente alla riduzione dello spessore dell'ossido di gate e al conseguente aumento della capacità dell'ossido C_{ox} . Tuttavia, la progressiva diminuzione di t_{ox} può generare una degradazione della qualità dell'ossido stesso e quindi un aumento della densità dei difetti (e quindi di trappole). Di conseguenza il parametro K_f potrebbe non diminuire nel modo aspettato.

Bisogna considerare, infine, il contributo di rumore Lorentziano che, come mostrato successivamente, viene osservato in alcuni dispositivi FinFET sotto misura. In un MOSFET di piccola area, i portatori non si trovano ad interagire con un numero elevato di trappole. Tuttavia, piccolo o grande che sia il numero di trappole, ciascuna di esse può determinare la cattura ed il rilascio dei portatori di carica secondo costanti di tempo ben definite, dipendenti dal livello energetico associato alla trappola e alla sua posizione, all'interfaccia o nell'ossido vicino ad essa, rispetto al canale conduttivo del dispositivo. Questo fenomeno può dare luogo ad una modulazione del flusso di portatori, provocando quindi una variazione Δi della corrente di canale. Il rumore che ne deriva, il cui studio può basarsi sulla teoria dei segnali telegrafici casuali (*Random Telegraph Signal, RTS*), è caratterizzato da una densità spettrale di potenza, che si aggiunge al contributo di rumore termico di canale ed al contributo di tipo $1/f$, data dall'espressione [41]:

$$\frac{di_L^2}{df} = \Delta i^2 \frac{4\tau_Z}{1 + \omega^2\tau_Z^2} \quad (3.17)$$

dove τ_Z è la costante di tempo associata alla trappola.

Secondo il modello proposto da McWorther, il rumore $1/f$ potrebbe trarre origine dalla sovrapposizione di un numero sufficientemente elevato di componenti lorentziane associate a trappole con opportuna distribuzione spaziale e conseguente opportuna distribuzione di tempi caratteristici, come si evidenzia in figura 3.5. In particolar modo, in un MOSFET il processo di cattura avviene per effetto tunneling dei portatori di carica dal semiconduttore alle trappole

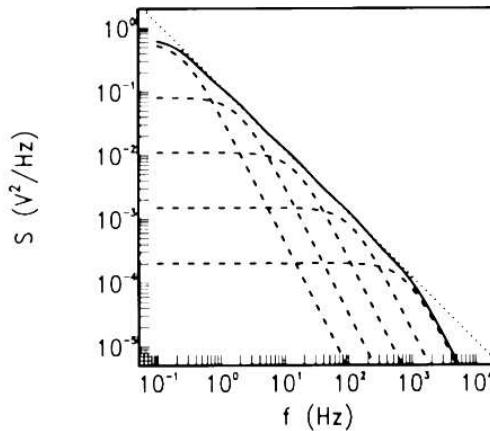


Figura 3.5: sovrapposizione di contributi lorentziani e origine del rumore flicker secondo McWorther.

localizzate all'interno dello strato di ossido ad una profondità d . La costante di tempo obbedisce alla legge:

$$\tau_Z = \tau_0 e^{\xi d} \quad (3.18)$$

dove τ_0 e ξ sono costanti dipendenti dal materiale. In questo caso la densità spettrale di rumore si ottiene sommando i contributi di trappole a diverse profondità (corrispondenti perciò a diverse costanti di tempo). Supponendo che la densità di trappole sia costante con la profondità è possibile dimostrare che la densità spettrale del rumore assume un'espressione con andamento di tipo *flicker*:

$$\frac{di_{1/f}^2}{df} = \frac{\Delta i^2 A N_t}{4\xi} \frac{1}{f}, \quad (3.19)$$

dove N_t rappresenta la densità di trappole e A l'area dove si ha intrappolamento.

3.2 Strumentazione per misure di rumore in dispositivi CMOS nanometrici

La caratterizzazione del rumore nei dispositivi a semiconduttore è di fondamentale importanza per capire i limiti di un dato processo tecnologico. Nel

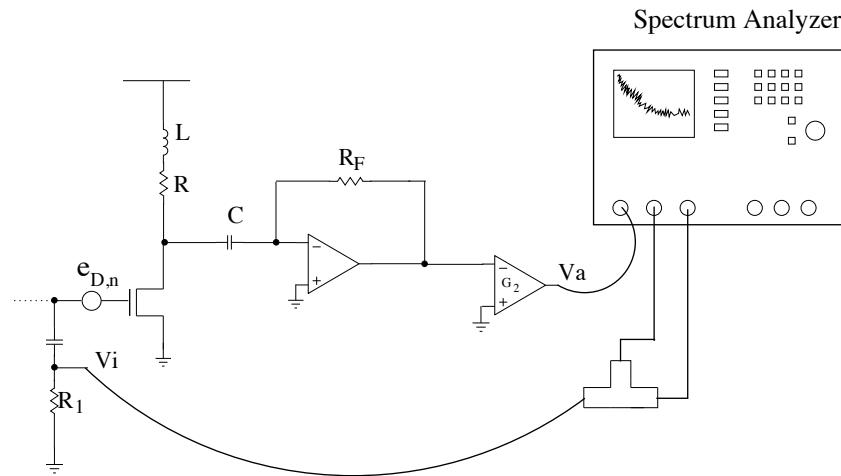


Figura 3.6: setup per la misura del rumore serie in un MOSFET basato su un amplificatore a transimpedenza.

dominio della frequenza la misura del rumore richiede un analizzatore di spettro, che, essendo costituito da circuiti elettronici, è intrinsecamente rumoroso. Di conseguenza si rende necessario l'impiego di ulteriori blocchi di interfaccia che permettano di amplificare selettivamente una delle sorgenti di rumore equivalente (serie o parallelo) in ingresso al DUT, in maniera tale da riportare all'ingresso dell'analizzatore di spettro, debitamente amplificato, il solo contributo di rumore desiderato. Il sistema deve inoltre consentire la polarizzazione del dispositivo sotto misura in diverse condizioni di lavoro, caratteristica necessaria se si desidera collaudare il DUT in condizioni il più possibile vicine a quelle di utilizzo reale, o se si intende effettuare una caratterizzazione più generale della tecnologia di appartenenza. A tutto questo si deve aggiungere la necessità che il blocco circuitale utilizzato per l'amplificazione del rumore stesso non fornisca un contributo di rumore significativo rispetto a quello del dispositivo sotto misura.

3.2.1 Setup di misura

In un transistore MOSFET, il contributo di rumore principale, ovvero il rumore nella corrente di drain, può essere modellizzato con un generatore di

tensione in serie al gate del DUT con densità spettrale di potenza:

$$\frac{de_{D,n}^2}{df} = \frac{4K_B T \Gamma}{g_m} \quad (3.20)$$

La misura effettiva prevede in realtà, come già accennato, che il rumore venga amplificato da un circuito di interfaccia prima di essere inviato all'ingresso dell'analizzatore di spettro. Al fine di poter estrarre, dalla misura effettuata, la densità spettrale di rumore riferito al gate del DUT, è necessario stabilire una procedura che consenta di compensare le possibili piccole variazioni nel guadagno, eventualmente anche in banda passante, del circuito di interfaccia stesso.

La figura 3.6 mostra il setup utilizzato per la misura della componente serie di rumore in un transistor FET, basato sull'utilizzo di un amplificatore a transimpedenza, trattato nel paragrafo 3.2.2. Questo stadio costituisce il primo blocco della catena di misura ed è seguito da un secondo stadio di guadagno G_2 che, in termini generali, può in realtà essere costituito a sua volta da diversi stadi di amplificazione. Il segnale in uscita è inviato all'analizzatore di spettro, che può operare anche come analizzatore di rete al fine di misurare la funzione di trasferimento tra il generatore equivalente di rumore in serie al gate del DUT, $e_{D,n}$, e l'ingresso dell'analizzatore, V_a .

La procedura di misura può essere riassunta come segue:

- si misura la risposta in frequenza ($T(j\omega)$) tra l'ingresso V_i e l'ingresso dell'analizzatore di spettro V_a . Questa misura viene effettuata facendo operare l'analizzatore di spettro come analizzatore di rete: esso invia un segnale sinusoidale all'ingresso V_i di ampiezza nota e frequenza variabile e misura l'ampiezza (e la fase) del segnale all'uscita della rete al variare della frequenza;
- si misura la densità spettrale di rumore in uscita $\frac{dV_{a,n}^2}{df}$;
- si ricava la densità spettrale di potenza del rumore del DUT riferito al suo gate come:

$$\frac{de_{D,n}^2}{df} = \frac{1}{|T(j\omega)|^2} \cdot \frac{dV_{a,n}^2}{df}. \quad (3.21)$$

Nello specifico, il sistema utilizzato come interfaccia tra DUT ed analizzatore di spettro, trattato in dettaglio nel paragrafo successivo, è realizzato su circuito stampato e viene mostrato schematicamente, in figura 3.7. Il chip, che ospita le strutture sotto misura FinFET, è alloggiato in un package che, a sua volta,

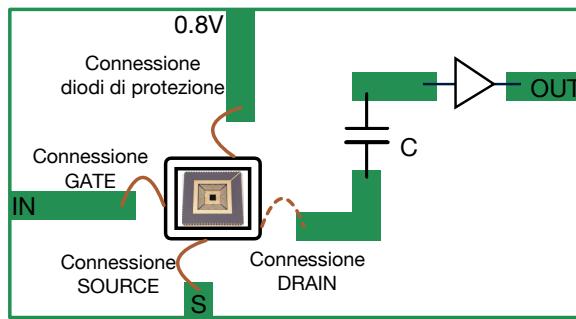


Figura 3.7: schema del sistema di misura realizzato su circuito stampato.

viene inserito in un socket montato sulla scheda di test. I dispositivi analizzati hanno ciascuno il proprio contatto di drain e presentano un contatto di source e di gate comune (due terminali, gate e source, per tutti i FinFET a canale P e due per i FinFET a canale N). Ipotizzando di voler misurare il rumore in un transistor a canale N, il drain del dispositivo da analizzare viene collegato all'ingresso dell'amplificatore a transimpedenza, attraverso una capacità ceramica di disaccoppiamento, anch'essa saldata sulla board di test e di valore $C \approx 10 \mu F$. La connessione a 0.8 V permette la polarizzazione dei diodi per la protezione dalle scariche elettrostatiche (ESD). Infatti, il contatto del chip con oggetti che, come il corpo umano, possono trovarsi a potenziali anche molto diversi rispetto a quelli presenti sul chip stesso, può danneggiare in modo irreversibile i dispositivi CMOS tramite perforazione del dielettrico tra gate e canale o tramite spostamento di carica nell'ossido, con conseguente variazione della tensione di soglia. L'applicazione di un'elevata tensione tra gate e substrato dà luogo a un campo elettrico molto forte nello strato di ossido, che ha spessore nanometrico, portando così al superamento della rigidità dielettrica o alla migrazione di cariche. Per questo motivo, all'ingresso dei pad vengono poste delle strutture di protezione (tipicamente diodi o transistori connessi a diodo) che impediscono che la tensione tra i terminali del dispositivo e massa e tra i terminali e l'alimentazione superi valori di sicurezza. Uno schema di principio di un possibile circuito di protezione viene riportato in figura 3.8, dove si utilizzano due diodi di clamping che entrano in conduzione quando la tensione di ingresso diventa più grande della tensione di alimentazione più una tensione di cut-in ($V_{DD} + V_\gamma$) o più piccola della massa meno una tensione di cut-in ($-V_\gamma$).

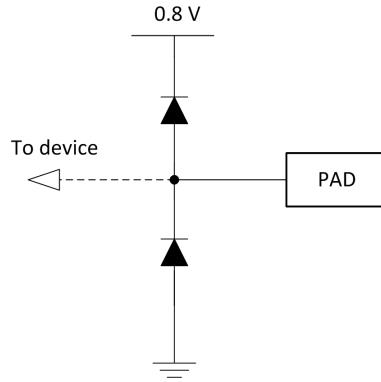


Figura 3.8: schema di una possibile struttura di protezione per le ESD.

Nonostante le contromisure adottate nel progetto del circuito di interfaccia, analizzate nel paragrafo 3.2.3, il livello di rumore all’uscita dell’amplificatore comprende contributi provenienti dal dispositivo sotto misura (desiderato) e contributi provenienti dalle sorgenti di rumore presenti nello schema di amplificazione. Per una corretta interpretazione delle misure di rumore è dunque necessario sottrarre quadraticamente al risultato della misura fornito dallo strumento il rumore di fondo dell’amplificatore. Con rumore di fondo si intende il valore che viene letto dall’analizzatore di spettro in assenza del DUT all’interno del circuito di amplificazione, ovvero quando il DUT non è alimentato.

3.2.2 Amplificatore a transimpedenza e rete di polarizzazione

Nel paragrafo 3.2 è stata evidenziata la necessità di disporre di un circuito in grado di amplificare la componente di rumore che si vuole misurare e che provveda alla polarizzazione del dispositivo stesso. Esistono differenti tipologie di amplificatori in grado di svolgere queste funzioni. Tuttavia, poiché il dispositivo sotto misura è un transistor MOS, generalmente caratterizzato da frequenze di corner di rumore (frequenza alla quale il rumore bianco eguaglia il rumore $1/f$) piuttosto elevate, anche superiori a 10 MHz, è necessario disporre di un amplificatore che sia in grado di fornire un’amplificazione su una banda di frequenza di circa 100 MHz. In questo modo lo strumento di misura consentirà di estrarre dallo spettro di rumore sia la componente bianca, associata al rumore termico di canale, sia la componente di tipo flicker.

In figura 3.9 viene mostrato lo schema a blocchi dell’amplificatore a transim-

pedenza e del sistema di misura utilizzato. Il rumore della corrente di drain

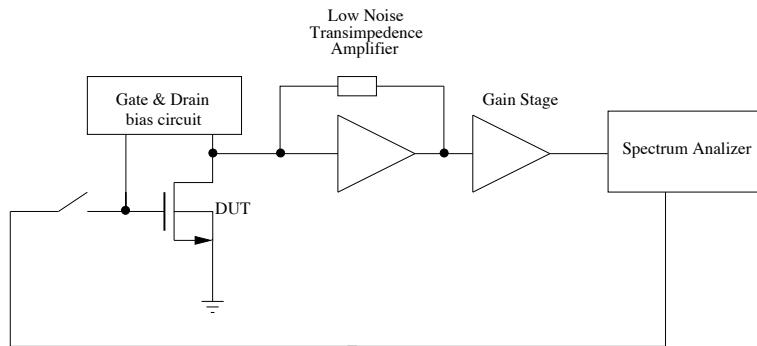


Figura 3.9: schema a blocchi dell'amplificatore e del sistema di misura utilizzato.

viene inizialmente amplificato e convertito in una tensione dall'amplificatore a trasimpedenza. Successivamente, la tensione di rumore risultante viene amplificata da uno stadio di guadagno e misurata dall'analizzatore di spettro.

Polarizzazione del DUT

Il circuito di polarizzazione del dispositivo sotto misura, il cui schema è riportato in figura 3.10, viene qui descritto per un N-MOS. Tuttavia uno schema del tutto analogo è stato utilizzato anche per la polarizzazione dei dispositivi a canale P. Il valore della corrente di drain I_D nel DUT viene fissato dalla tensione di gate, ottenuta facendo variare la tensione V_G ai capi di un partitore resistivo realizzato con le resistenze R_2 ed R_3 , di ugual valore. La resistenza R_4 e la capacità C_{in} costituiscono invece un filtro passa-basso che ha lo scopo di filtrare il rumore delle resistenze di polarizzazione nella banda di interesse. Per la scelta dei componenti di questo filtro, si osserva che da un lato si ha la necessità di minimizzare la frequenza di taglio, in modo da eliminare i contributi di rumore delle resistenze di polarizzazione, dall'altro una costante di tempo del filtro troppo elevata, all'atto dell'accensione del circuito, manterrebbe il drain del dispositivo a tensioni superiori a quelle consentite dalla tecnologia per un tempo troppo lungo. Conseguentemente, per R_4 è stato scelto un valore di $10 \text{ k}\Omega$ e per C_{in} un valore pari a $2.2 \mu\text{F}$. Si ottiene in questo modo un filtro passa-basso caratterizzato da una frequenza di taglio ed un tempo di salita

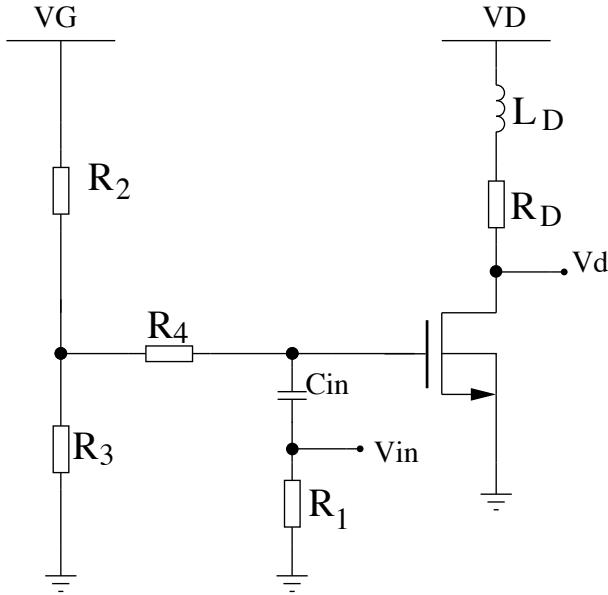


Figura 3.10: circuito di polarizzazione del DUT.

dati rispettivamente da:

$$f_{HLP} = \frac{1}{2\pi R_4 C_{in}} = 13.8 \text{ Hz}; \quad t_{SLP} = \frac{0.35}{f_{HLP}} = 25 \text{ msec.} \quad (3.22)$$

Una volta stabilito il valore della corrente, è possibile fissare la tensione di drain V_d attraverso la resistenza R_D , facendo variare questa volta la tensione V_D . La presenza dell'induttanza L_D ha il solo scopo di filtrare il rumore della resistenza R_D nella banda di misura. Considerando infatti il circuito semplificato di figura 3.11 è possibile ricavare la funzione di trasferimento per la sorgente di rumore associata alla resistenza, che risulta essere di tipo passa-basso:

$$\frac{V_{out}}{V_{n,L_D}} = \frac{R_F}{R_D} \frac{1}{1 + s \frac{L_D}{R_D}}. \quad (3.23)$$

La frequenza di taglio del filtro, f_T , è data da:

$$f_T = \frac{1}{2\pi \frac{L_D}{R_D}}. \quad (3.24)$$

La resistenza R_1 , di valore pari a 50Ω , ha lo scopo di fornire il necessario adattamento di impedenza per il cavo con impedenza caratteristica di 50Ω ,

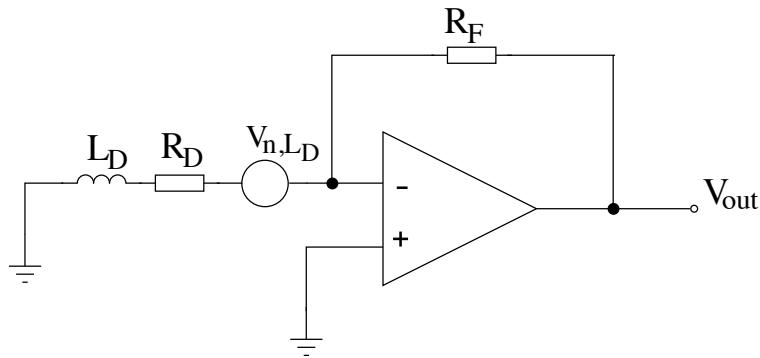


Figura 3.11: schema semplificato per il calcolo del contributo di rumore della resistenza R_D .

che connette l'uscita dell'analizzatore all'ingresso del circuito di misura.

Lo stadio a transimpedenza è realizzato con un amplificatore operazionale commerciale (LMH6624) che viene alimentato tra +5 V e -5 V. Nel seguito si analizzano le principali caratteristiche dell'amplificatore a transimpedenza utilizzato per le misure di rumore.

Banda e guadagno dell'amplificatore

Per l'analisi delle caratteristiche di guadagno e di larghezza di banda dell'amplificatore è utile considerare lo schema semplificato rappresentato in figura 3.12, dove l'amplificatore a transimpedenza viene rappresentato mediante un modello a singolo polo, con funzione di trasferimento del tipo:

$$A(s) = \frac{A_0}{1 + s\tau}. \quad (3.25)$$

R_F e C_F rappresentano, rispettivamente, la resistenza e la capacità di reazione; quest'ultima può essere anche solo costituita dalla capacità parassita presente tra ingresso ed uscita dell'amplificatore. La capacità C_D costituisce, invece, la somma della capacità presente tra drain e source del DUT, di quella di ingresso dell'amplificatore operazionale e di quella parassita presente al terminale di drain del DUT (dipendente dalle caratteristiche del *package*, dello zoccolo e della scheda stampata). Infine, il generatore di tensione $e_{D,n}$ rappresenta il generatore equivalente di rumore serie del FET sotto misura.

Con riferimento alla figura 3.12 si ottiene dunque una funzione di trasferimento di tipo passa-basso con due poli complessi coniugati e due zeri di trasmissione

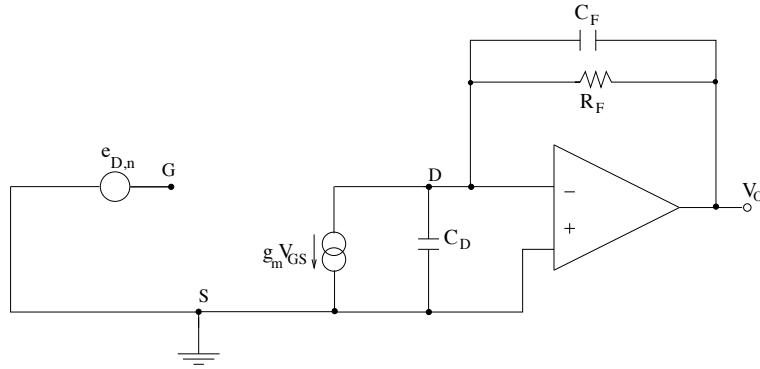


Figura 3.12: schema semplificato del amplificatore e circuito equivalente di piccolo segnale del DUT. Nello schema è incluso anche il generatore equivalente serie di rumore, $e_{D,n}$.

all'infinito del tipo:

$$T(s) = \frac{V_0(s)}{e_{D,n}(s)} = \frac{A_f}{s^2 + s\frac{\omega_0}{Q} + \omega_0^2} \quad (3.26)$$

con

$$A_f = \frac{g_m A_0}{\tau_0(C_D + C_F)} \approx \frac{g_m A_0}{\tau_0 C_D} = \frac{2\pi GBP}{C_D/g_m} \quad (3.27)$$

$$\omega_0^2 = \frac{1 + A_0}{\tau_0 R_F (C_D + C_F)} \approx \frac{A_0}{\tau_0} \cdot \frac{1}{R_F (C_D + C_F)} \approx \frac{2\pi GBP}{R_F C_D} \quad (3.28)$$

$$\frac{\omega_0}{Q} = \frac{\tau_0 + R_F [C_D + C_F (1 + A_0)]}{\tau_0 R_F (C_D + C_F)} \approx \frac{C_F A_0}{\tau_0 C_D} = \frac{2\pi GBP}{C_D/C_F} \quad (3.29)$$

dove $GBP = A_0/2\pi\tau_0$ rappresenta il prodotto banda-guadagno dell'amplificatore operazionale e Q il fattore di merito. La figura 3.13 chiarisce il significato dei termini utilizzati nell'espressione (3.26).

In base alle precedenti espressioni, è possibile risalire al valore del guadagno in continua dato da:

$$T(0) = A_{f0} = \frac{A_f}{\omega_0^2} \approx g_m R_F. \quad (3.30)$$

La (3.26) evidenzia che il modulo della risposta in frequenza $T(j\omega)$ può presentare un picco di risonanza dipendente dal fattore di qualità Q del polo, caratterizzato da un'ampiezza:

$$A_{f,max} = \frac{A_f Q}{\omega^2 \sqrt{1 - (1/4Q^2)}} \quad (3.31)$$

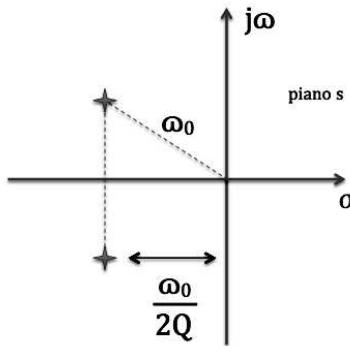


Figura 3.13: definizione di ω_0 e Q per una coppia di poli complessi coniugati.

e localizzato alla frequenza angolare ω_{max} , data da:

$$\omega_{max} = \omega_0 \sqrt{1 - (1/2Q^2)}. \quad (3.32)$$

Una risposta in frequenza che presenta un picco di risonanza pronunciato può comportare non linearità nella risposta del circuito ed errori di misura. Per evitare questo si tende ad avere una funzione di trasferimento che sia il più possibile costante, o massimamente piatta, in banda passante, agendo sul fattore Q del polo mediante l'introduzione di una capacità in parallelo a R_F . Per una risposta massimamente piatta si deve soddisfare la seguente condizione:

$$A_{f,max} = A_{f0} \Rightarrow Q = 1/\sqrt{2}. \quad (3.33)$$

Di conseguenza, dalle equazioni (3.28) e (3.29), in accordo con la (3.33), si ottiene per il fattore di qualità Q l'espressione:

$$Q = \sqrt{\frac{\omega_0^2}{(\omega_0/Q)^2}} \approx \frac{1}{C_F} \cdot \sqrt{\frac{C_D}{2\pi \cdot GBP \cdot R_F}} = 1/\sqrt{2}. \quad (3.34)$$

Queste relazioni forniscono dei criteri per la scelta dei valori dei componenti che costituiscono la rete di reazione e dell'amplificatore operazione, del quale, oltre al prodotto banda-guadagno, è importante valutare la capacità di ingresso, che influenza il valore di C_D e con esso quello della banda disponibile.

In particolare R_F , ha un valore di $10 k\Omega$, al fine di garantire un buon compromesso tra guadagno di transimpedenza e rumore, mentre il valore di C_F , che costituisce la capacità parassita in parallelo a R_F , è dell'ordine di qualche frazione di pF .

Input Referred Voltage Noise $(\sqrt{S_{v,OP}(\omega)})$	Input Referred Current Noise $(\sqrt{S_{i,OP}(\omega)})$	Gain Bandwidth (GBP)	Input Capacitance (C_D)
0.92 nV/ \sqrt{Hz}	2.3 pA/ \sqrt{Hz}	1.5 GHz	$\approx 3 \text{ pF}$

Tabella 3.1: caratteristiche dell’amplificatore LMH6624.

L’amplificatore commerciale LMH6624, con cui è realizzato l’amplificatore a transimpedenza, consente di eseguire misure fino ad una frequenza di circa 100 MHz. I suoi parametri principali sono riportati nella tabella 3.1. La scelta di questo dispositivo è giustificata in particolare dalle sue proprietà in termini di prodotto banda-guadagno e di rumore equivalente in ingresso.

Bisogna infine considerare il fatto che il rumore intrinseco dell’analizzatore di spettro, un modello HP4195A prodotto da HP, contribuisce in modo significativo al rumore totale su tutta la banda di frequenze di misura. La sua densità spettrale, riferita all’ingresso dello strumento, risulta essere $S_{HP} = 10 \text{ nV}/\sqrt{Hz}$; se riportata all’ingresso dell’amplificatore a transimpedenza, vale $S_{IN,HP} = 2 \text{ pA}/\sqrt{Hz}$. Al fine di rendere trascurabile questo contributo di rumore indesiderato è possibile introdurre uno stadio di guadagno aggiuntivo tra l’amplificatore a transimpedenza e l’analizzatore di spettro. Tale stadio, realizzato mediante l’amplificatore commerciale HP8447C, fornisce un guadagno di 30 dB sulla banda da 30 a 300 MHz. La risposta in frequenza $T(j\omega)$ del sistema di misura, ricavata sperimentalmente, viene mostrata in figura 3.14.

3.2.3 Analisi delle sorgenti di rumore

Al fine di eseguire una misura affidabile del rumore nel DUT, il sistema di misura, ed in particolare lo stadio a transimpedenza, deve contribuire in modo trascurabile al rumore totale in uscita. Il rumore dei componenti attivi e passivi presenti nel circuito di interfaccia è minimizzato mediante la scelta opportuna dei valori dei componenti critici o adottando accorgimenti particolari, come visto nella sezione in cui si è discusso della rete di polarizzazione del DUT.

Nella banda in cui vengono condotte le misure, gli unici componenti critici dal punto di vista del contributo di rumore in uscita sono: la resistenza di retroazione R_F , la resistenza di adattamento R_1 e l’amplificatore operazionale. La resistenza R_1 nella procedura di misura viene posta in parallelo ad una terminazione da 50Ω e di conseguenza il contributo di rumore equivalente è

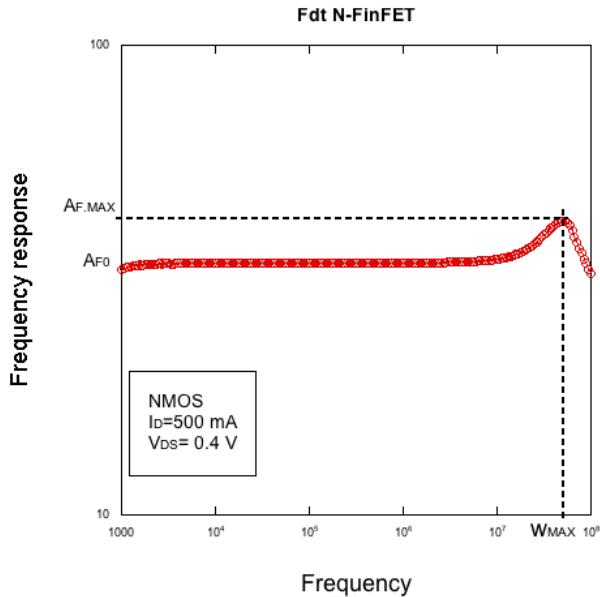


Figura 3.14: risposta in frequenza del sistema di misura.

quello di una resistenza di valore pari a 25Ω . Questa contribuisce quindi con una densità spettrale di rumore in serie al gate del DUT pari a $4K_B T(R_1/2)$ che viene sottratta quadraticamente al rumore misurato.

Con riferimento alla figura 3.15, il rumore nella resistenza R_F è rappresentato da un generatore di corrente i_{RF} posto in parallelo alla resistenza stessa con densità spettrale uguale a:

$$S_{RF} = \frac{4K_B T}{R_F}. \quad (3.35)$$

Il rumore dell'amplificatore operazionale è descritto da un contributo serie e da un contributo parallelo mediante i generatori v_{OP} e i_{OP} , con densità spettrale di potenza rispettivamente dati da $S_{v,OP}$ e $S_{i,OP}$. Infine il contributo di rumore del DUT viene rappresentato mediante il generatore di corrente parallelo i_d , caratterizzato dalla densità spettrale di rumore S_{i_d} . Valutare il contributo delle sorgenti di rumore descritte, esclusa quella associata al DUT, significa fornire una stima del rumore di fondo dello strumento. Questo determina la sensibilità con cui è possibile eseguire la misura: il valore del rumore misurato in uscita può essere considerato attendibile solo se è molto maggiore di quello di fondo misurato nelle stesse condizioni operative. Dallo schema riportato in figura 3.15 è possibile ricavare l'espressione analitica della densità spettrale di

rumore in uscita, data da:

$$S_{out} = [S_{RF} + S_{i,OP}] \cdot |T_1(j\omega)|^2 + S_{v,OP} \cdot |T_2(j\omega)|^2 \quad (3.36)$$

dove $T_1(j\omega)$ e $T_2(j\omega)$ sono rispettivamente le funzioni di trasferimento associate ai contributi di rumore di tipo parallelo e di tipo serie. La funzione di trasferimento $T_1(s)$ è data da:

$$T_1(s) = \frac{v_{OUT}}{i_{OP}} = \frac{R_F}{1 + sR_F C_F} \quad (3.37)$$

mentre, per $T_2(s)$ si ricava l'espressione:

$$T_2(s) = \frac{v_{OUT}}{v_{OP}} = \frac{1 + sR_F(C_D + C_F)}{1 + sR_F C_F}. \quad (3.38)$$

Al fine di valutare l'impatto delle sorgenti di rumore dello stadio a transimpedenza sulle prestazioni del sistema di misura, è utile riferire i vari contributi di rumore all'ingresso dell'amplificatore in termini di rumore parallelo. La densità spettrale di potenza $S_{TOT_{IN}}$ associata al rumore di fondo consiste dunque nella somma delle singole componenti di rumore riportate all'ingresso dell'amplificatore, ovvero:

$$S_{TOT_{IN}} = S_{IN_{//}} + S_{IN_{SERIE}} \quad (3.39)$$

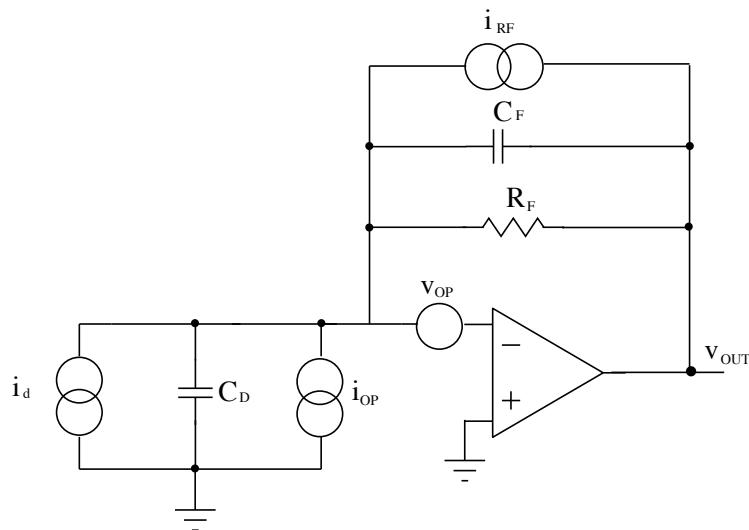


Figura 3.15: stadio a transimpedenza con generatori di rumore.

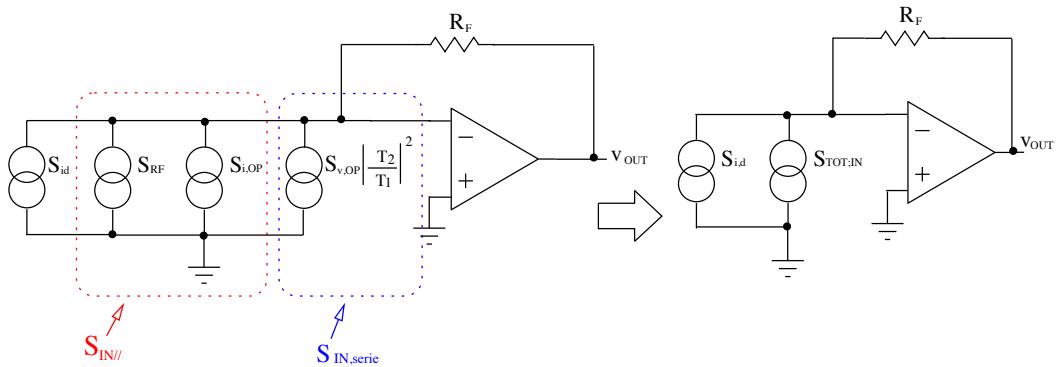


Figura 3.16: rappresentazione dei vari contributi di rumore all’ingresso dell’amplificatore a transimpedenza.

dove $S_{IN,//}$ rappresenta la densità spettrale di potenza del rumore in ingresso dovuta alle sorgenti di tipo parallelo e $S_{IN,SERIE}$ quella dovuta alle sorgenti di tipo serie. Queste sorgenti sono rappresentate schematicamente in figura 3.16. La componente di tipo parallelo è data dalla somma dei singoli contributi dovuti alla resistenza di reazione R_F e alla componente $S_{i,OP}$, ovvero:

$$S_{IN,//} = S_{i,OP} + S_{RF}. \quad (3.40)$$

La componente serie, dovuta alla sorgente $S_{v,OP}$, ha la seguente espressione:

$$S_{IN,SERIE} = S_{v,OP} \cdot \left| \frac{T_2(j\omega)}{T_1(j\omega)} \right|^2. \quad (3.41)$$

Il contributo di queste sorgenti di rumore viene tenuto in considerazione mediante la misura sistematica del rumore di fondo, il quale viene sottratto quadraticamente al rumore misurato al fine di ottenere risultati indipendenti dal sistema di acquisizione. Tutti i dati che vengono presentati nel paragrafo successivo sono pertanto da intendere come depurati dal contributo di rumore proveniente dal sistema di misura.

La figura 3.17 mostra il confronto tra la densità spettrale di rumore di un FinFET a canale N e il rumore di fondo del sistema di misura per una corrente di drain pari a $25 \mu\text{A}$ e $500 \mu\text{A}$. Come si può osservare, il peso del rumore di fondo è più significativo alle alte frequenze e per basse correnti di drain.

Per misure in alta frequenza (superiori a 100 MHz) il contributo dominante del rumore di fondo deriva dalla componente serie all’ingresso dell’amplificatore

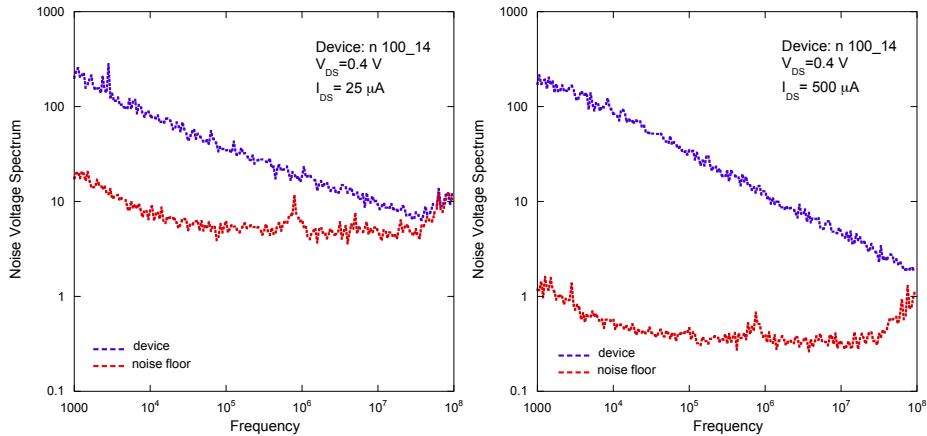


Figura 3.17: radice della densità spettrale di rumore serie in un FinFET a canale N con $W/L = 100 \mu\text{m}/14\text{nm}$ in tecnologia 14 nm (in blu) e rumore di fondo del sistema di misura (in rosso).

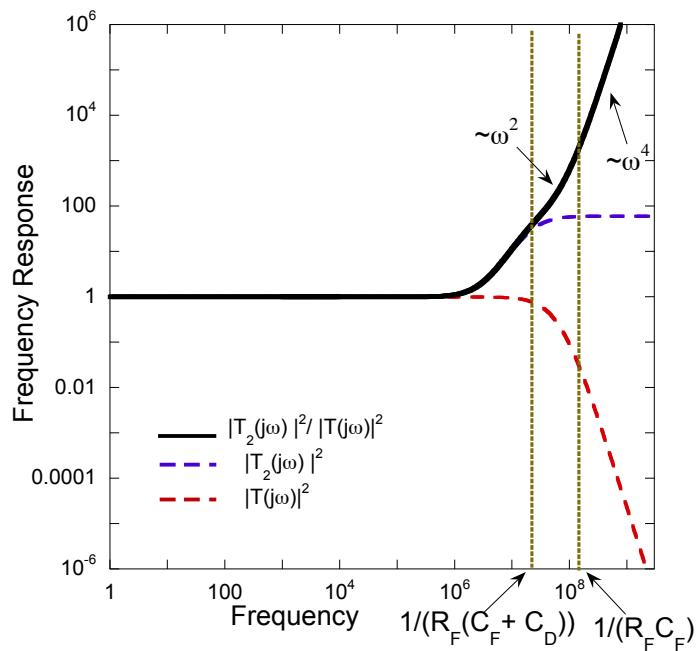


Figura 3.18: andamento di $|T_2(j\omega)|^2$, $|T(j\omega)|^2$ e del loro rapporto.

a transimpedenza. Facendo riferimento alle equazioni (3.26) e (3.38) e assumendo, a titolo di esempio, valori tipici per i vari parametri in gioco, ovvero $R_F = 10 \text{ k}\Omega$, $C_D = 5 \text{ pF}$, $C_F = 0.5 \text{ pF}$, $GBP = 1.5 \text{ GHz}$ e $g_m = 100 \mu\text{A/V}$, si ricava l'andamento della risposta in frequenza $|T_2(j\omega)|^2$ (curva a tratto blu) e di $|T(j\omega)|^2$ (curva a tratto rosso) mostrati in figura 3.18. Il quadrato del rapporto tra $|T_2(j\omega)|$ e $|T(j\omega)|$ rappresenta il termine che moltiplica $S_{v,OP}$ nella (3.41), che a sua volta fornisce il contributo del rumore serie del sistema riferito all'ingresso dell'amplificatore a transimpedenza. L'andamento di $|T_2(j\omega)|^2 / |T(j\omega)|^2$ è pure riportato in figura 3.18, dalla quale si nota che il termine cresce:

- come ω^2 per $\frac{1}{R_F \cdot (C_D + C_F)} < \omega < \frac{1}{R_F C_F}$;
- come ω^4 per $\omega > \frac{1}{R_F C_F}$.

In alta frequenza è infine necessario tener conto dei contributi parassiti legati alle piste metalliche sulla scheda stampata, ai cavi di collegamento ed allo zoccolo in cui il chip dei dispositivi sotto misura viene innestato.

3.3 Misure di rumore

In questo paragrafo vengono presentate le misure di densità spettrale di rumore serie nei dispositivi FinFET. Tale grandezza viene direttamente calcolata mediante un programma che, in maniera semiautomatica, dopo la determinazione della funzione di trasferimento $T(j\omega)$ tra l'ingresso del sistema e l'ingresso dell'analizzatore di spettro, riferisce la densità spettrale di potenza del rumore misurato (che include contributi di rumore termico di canale e di rumore *flicker*) all'ingresso del DUT in accordo con l'equazione (3.21). L'acquisizione dei dati avviene tramite un programma sviluppato in ambiente Labview.

3.3.1 Misure di densità spettrale di rumore

Le misure di rumore sono state condotte nel range di frequenza 1 kHz - 100 MHz. I dispositivi sotto misura sono stati polarizzati con una tensione $V_{DS} = 0.4 \text{ V}$ per i dispositivi a canale N, $V_{DS} = -0.4 \text{ V}$ per i dispositivi a canale P e con una tensione bulk-substrato $V_{BS} = 0 \text{ V}$. Al fine di valutare la dipendenza del rumore dalle condizioni di lavoro, i FinFET sono stati polarizzati con differenti valori di corrente di drain I_D pari a $25 \mu\text{A}$, $50 \mu\text{A}$, $100 \mu\text{A}$, $250 \mu\text{A}$, $500 \mu\text{A}$. I grafici seguenti rappresentano esempi di tipici risultati sperimentali. In particolar modo, le figure dalla 3.19 alla 3.24 mostrano la tensione di rumore serie (ovvero la radice della densità spettrale di rumore serie) di

alcuni dispositivi FinFET per differenti valori della corrente di drain. Nella regione dove il rumore *flicker* risulta dominante non si osservano significative differenze al variare della corrente di polarizzazione, in accordo col modello proposto da Mc Worther per descrivere l'origine del rumore 1/f nei MOSFET. Il rumore bianco, invece, decresce con l'aumento della corrente di drain I_D , come atteso in base all'equazione (3.10).

Infine, in due dei quattro dispositivi NMOS con $W/L = 600 \mu m/18 nm$, si evidenzia, a basse frequenze, un contributo di rumore in eccesso, forse imputabile a termini di tipo Lorentziano. Le figure relative a questi dispositivi sono mostrate in figura 3.25.

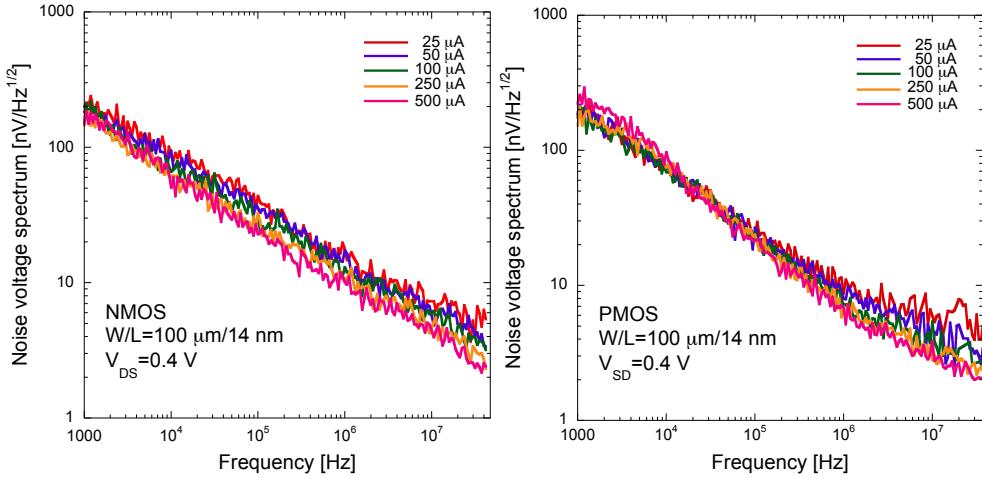


Figura 3.19: tensione di rumore serie in un FinFET a canale N (a sinistra) ed uno a canale P (a destra) con $W/L=100 \mu\text{m}/14 \text{ nm}$.

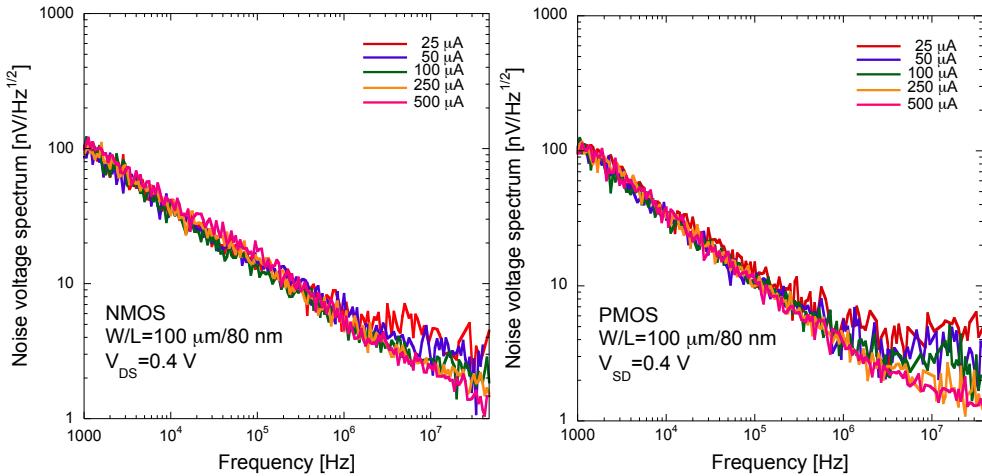


Figura 3.20: tensione di rumore serie in un FinFET a canale N (a sinistra) ed uno a canale P (a destra) con $W/L=100 \mu\text{m}/80 \text{ nm}$.

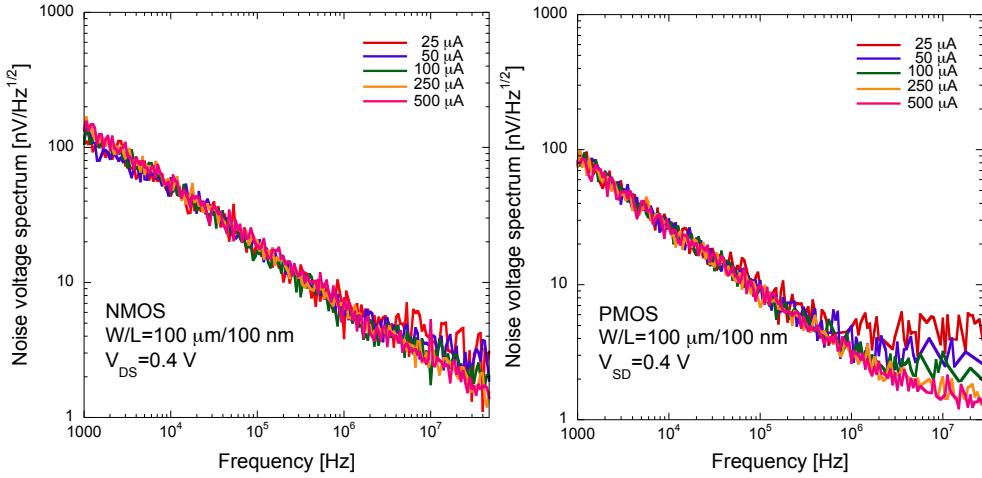


Figura 3.21: tensione di rumore serie in un FinFET a canale N (a sinistra) ed uno a canale P (a destra) con $W/L=100 \mu\text{m}/100 \text{ nm}$.

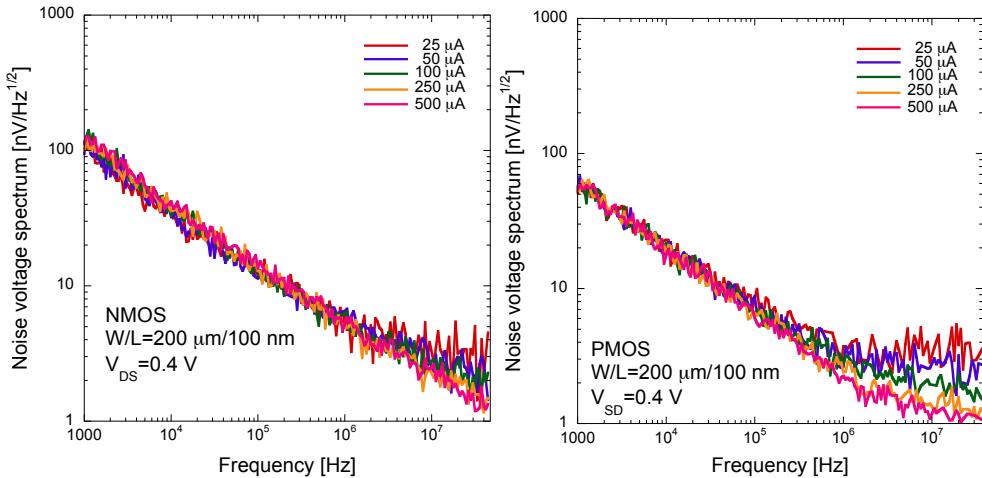


Figura 3.22: tensione di rumore serie in un FinFET a canale N (a sinistra) ed uno a canale P (a destra) con $W/L=200 \mu\text{m}/100 \text{ nm}$.

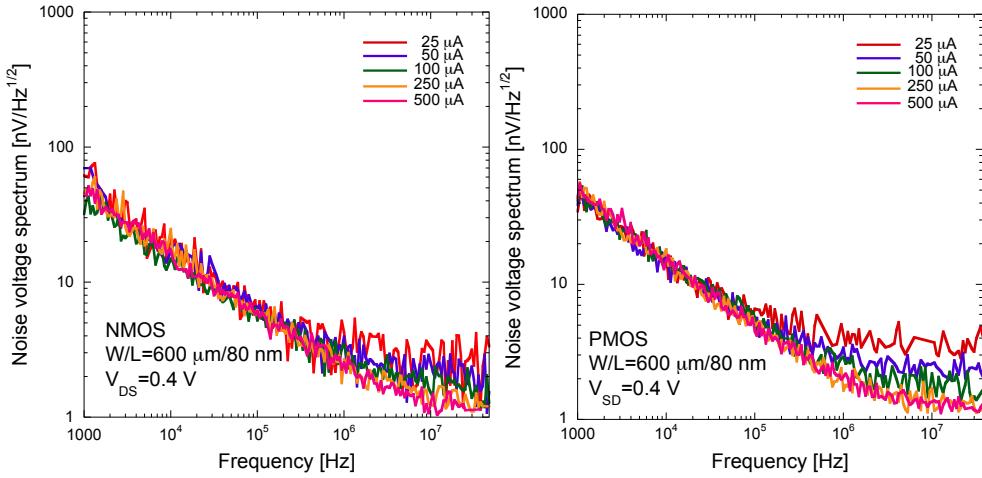


Figura 3.23: tensione di rumore serie in un FinFET a canale N (a sinistra) ed uno a canale P (a destra) con $W/L=600 \mu\text{m}/80 \text{ nm}$.

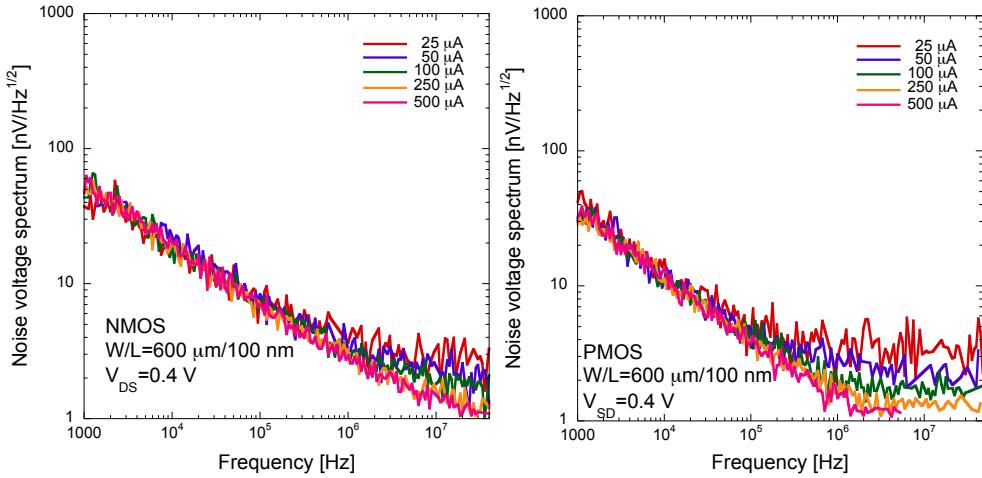


Figura 3.24: tensione di rumore serie in un FinFET a canale N (a sinistra) ed uno a canale P (a destra) con $W/L=600 \mu\text{m}/100 \text{ nm}$.

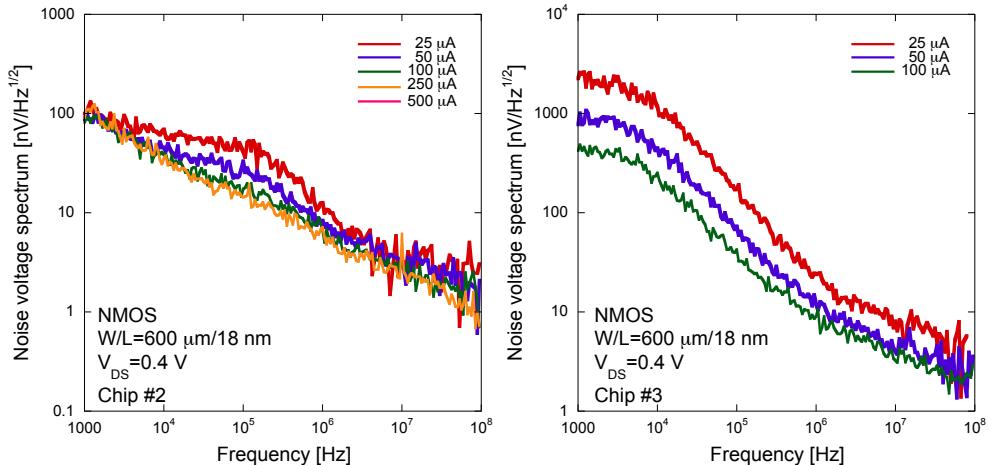


Figura 3.25: tensione di rumore serie in un FinFET a canale N con $W/L=600 \mu\text{m}/18 \text{ nm}$, con contributi d rumore in eccesso in bassa frequenza.

3.3.2 Dipendenza del rumore dalle dimensioni del gate

La figura 3.26 mostra la tensione di rumore in funzione della lunghezza di canale L per dispositivi NMOS (a sinistra) e PMOS (a destra) caratterizzati da una larghezza di canale W pari a $100 \mu\text{m}$ e corrente di drain $I_D = 25 \mu\text{A}$. Poiché i FET sono polarizzati in debole inversione, il rumore bianco è indipendente dalla lunghezza di gate, mentre a basse frequenze il rumore $1/f$ aumenta al diminuire di L , come suggerito dalla (3.15).

La figura 3.27 mostra la dipendenza della tensione di rumore serie dalla larghezza di gate W . In accordo con l'equazione (3.15), il rumore *flicker* aumenta al diminuire di W , mentre il rumore bianco, analogamente a quanto detto prima, non evidenzia una dipendenza significativa dalla geometria del canale.

Infine, le figure 3.28 e 3.29 mostrano il parametro A_f , definito come:

$$A_f = \frac{K_f}{C_{ox}WL}, \quad (3.42)$$

rispettivamente in funzione della lunghezza di gate, con larghezza W fissa, ed in funzione della larghezza di gate, con lunghezza L fissa. Queste figure evidenziano, come atteso, che il parametro A_f decresce all'aumentare della geometria di gate.

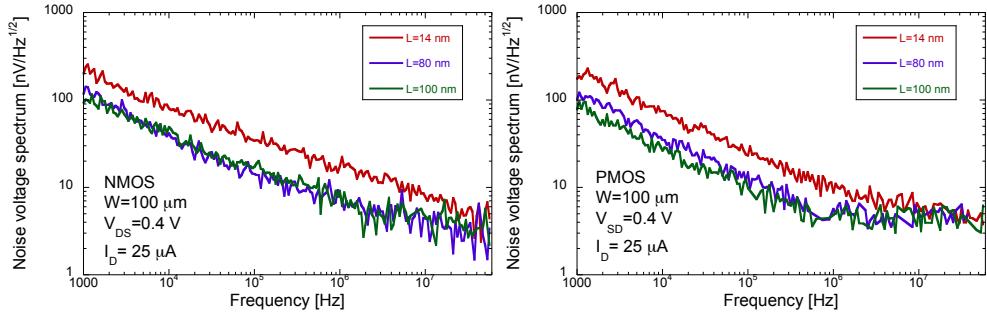


Figura 3.26: tensione di rumore serie in un FinFET a canale N (a sinistra) ed uno a canale P (a destra) con $W=100 \mu\text{m}$ e differenti valori di L .

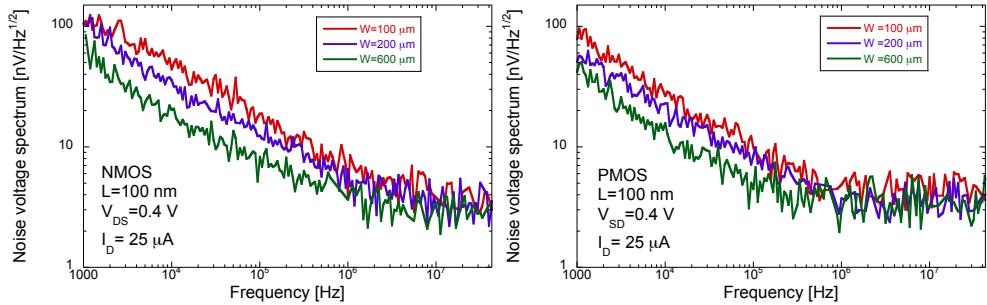


Figura 3.27: tensione di rumore serie di un FinFET a canale N (a sinistra) e a canale P (a destra) con $L=100 \text{ nm}$ e differenti valori di W .

3.3.3 Analisi dei parametri di rumore

Nel paragrafo seguente si discutono i principali parametri di rumore estratti dalle misure di densità spettrale.

Analisi del contributo di rumore bianco

Il rumore bianco viene qui valutato in termini di resistenza equivalente di rumore termico di canale R_{eq} , secondo la definizione data da (3.12). La figura 3.30 mostra i valori di R_{eq} in funzione di $n_{sub}\gamma/g_m$ estratti dai dispositivi a canale N e a canale P con differenti geometrie di gate. Per il coefficiente n_{sub} sono stati utilizzati i valori riportati in tabella 2.2, mentre il coefficiente γ è stato calcolato dalla (3.11) per i diversi valori di I_D e con i valori di I_Z^* riportati in tabella 2.2. In accordo con (3.12), il coefficiente di rumore bianco

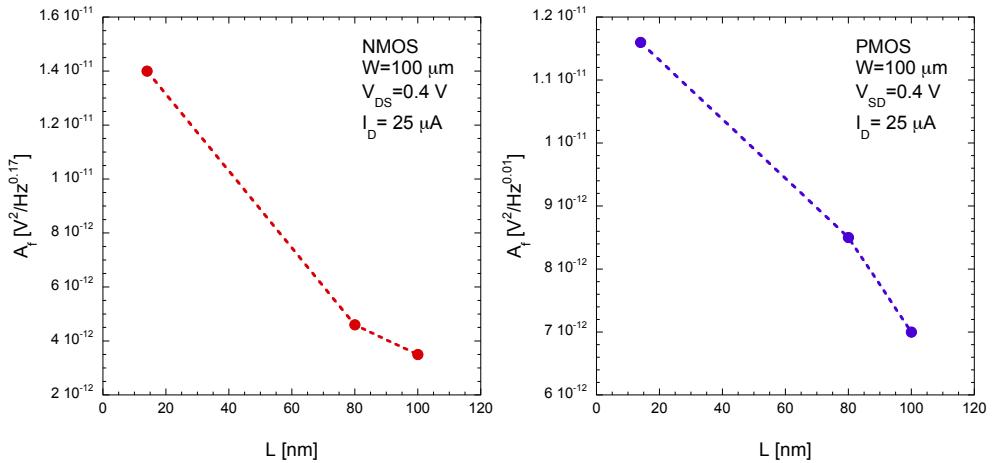


Figura 3.28: parametro A_f in un FinFET a canale N (a sinistra) ed uno a canale P (a destra) con $W=100$ nm e differenti valori di L .

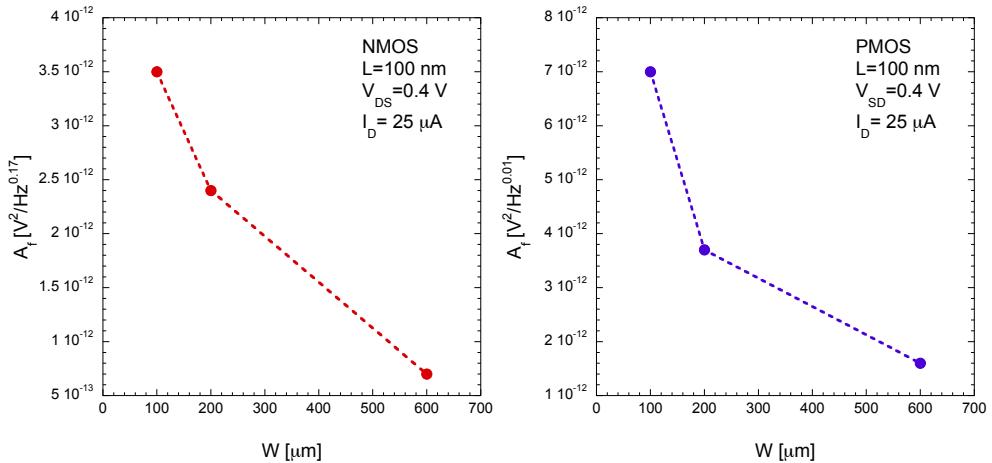


Figura 3.29: parametro A_f in un FinFET a canale N (a sinistra) ed uno a canale P (a destra) con $L=100$ nm e differenti valori di W .

α_w può essere determinato valutando la pendenza della rette interpolante come mostrato in figura 3.30, dove si osserva che i valori di α_w sono molto vicini all'unità per entrambe le polarità del dispositivo. Questo significa che non possono essere rilevati significativi effetti di canale corto nella regione operativa considerata e per i dispositivi caratterizzati.

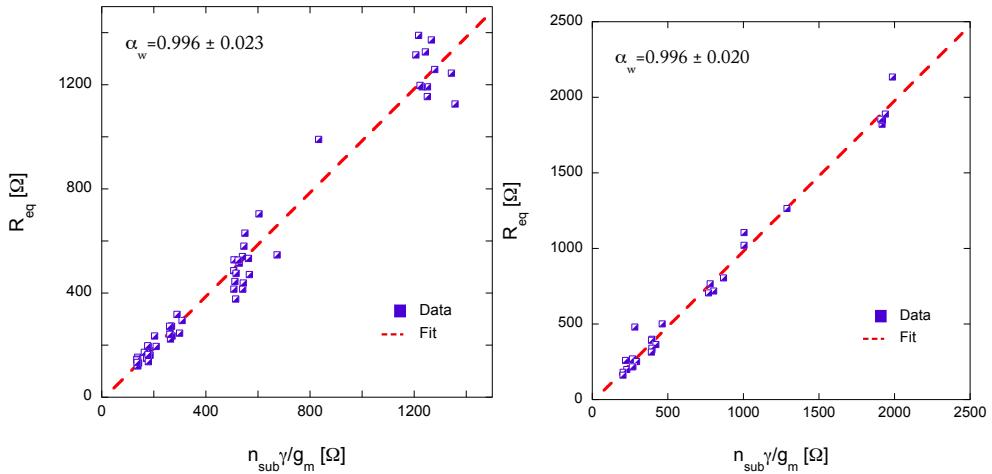


Figura 3.30: resistenza equivalente di rumore termico di canale R_{eq} estratta per dispositivi NMOS (a sinistra) e PMOS (a destra) con varie geometrie di gate, in funzione di $n_{sub}\gamma/g_m$.

Analisi del rumore *flicker*

L'analisi dei risultati sperimentali ottenuti in termini di densità spettrale di potenza di rumore serie, mostra che il coefficiente relativo alla pendenza del rumore $1/f$, α_f , è minore di 1 nei dispositivi NMOS e vicino all'unità nei dispositivi PMOS. È stato inoltre trovato che α_f non mostra una significativa dipendenza dalla geometria del canale (lunghezza e larghezza) o dalla corrente di drain I_D . L'istogramma in figura 3.31 evidenzia una dispersione relativamente ridotta del coefficiente α_f per i transistor testati, mentre la figura 3.32 confronta i valori di α_f ottenuti per la tecnologia studiata con quelli ricavati per alcuni processi CMOS appartenenti a precedenti generazioni.

Il valore di K_f/C_{ox} , in accordo con la (3.15), è stato estratto dalla densità spettrale di rumore utilizzando il valore medio del coefficiente α_f mostrato in figura 3.31. Il rapporto K_f/C_{ox} non permette un confronto diretto tra dispositivi PMOS ed NMOS o con vecchie tecnologie, in termini di prestazioni di rumore, a causa dei differenti valori di α_f . Il parametro K_f è infatti una misura dell' energia (espressa dunque in Joule) del rumore di tipo $1/f$ solo se $\alpha_f = 1$. Nel caso in cui il coefficiente α_f sia diverso dall'unità, come nella tecnologia analizzata (ma anche nelle vecchie tecnologie CMOS), la definizione circa l'energia del rumore *flicker* deve tenere conto della dipendenza dalla

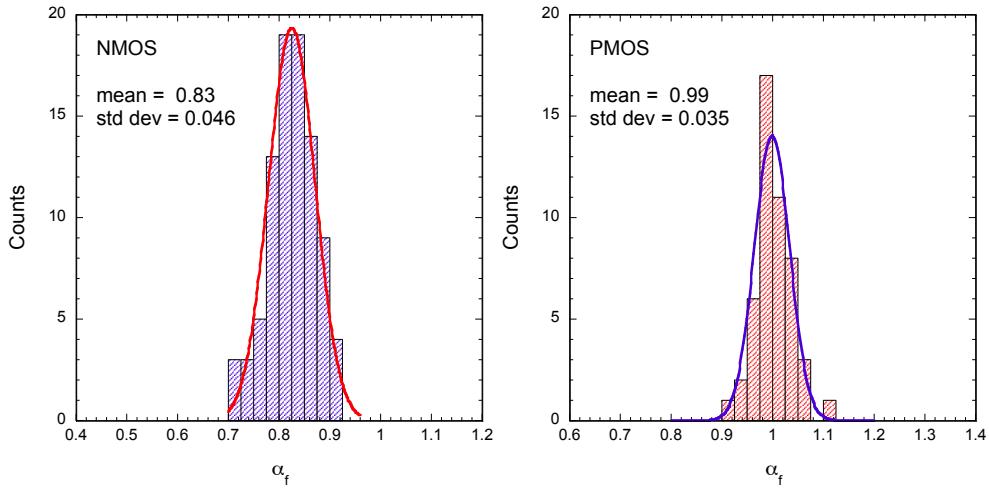


Figura 3.31: pendenza del contributo di rumore *flicker*, α_f , per i dispositivi NMOS e PMOS appartenenti alla tecnologia FinFET .

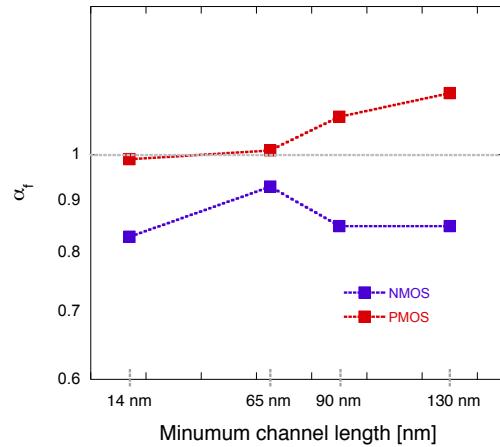


Figura 3.32: coefficiente α_f , nei dispositivi FinFET confrontato con il valore estratto da processi CMOS appartenenti a precedenti nodi tecnologici.

frequenza. A tale scopo può essere definito il parametro $M_f(f)$:

$$M_f(f) = \frac{K_f}{C_{ox}} f^{1-\alpha_f}. \quad (3.43)$$

Il valore calcolato di $M_f(f)$ a frequenza $f=1$ kHz è di $16.8 \cdot 10^{-21}$ e

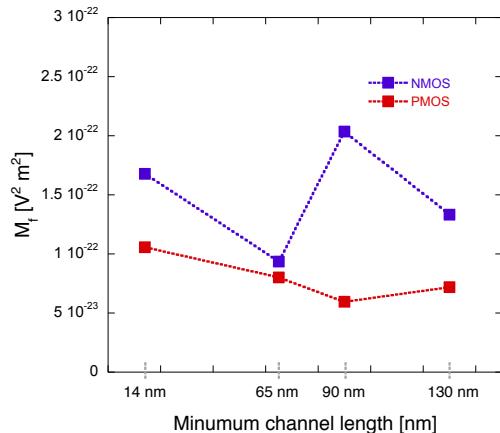


Figura 3.33: parametro M_f della tecnologia esaminata a confronto con il valore estratto da processi CMOS appartenenti a precedenti nodi tecnologici.

$10.6 \cdot 10^{-21} V^2 m^2$ rispettivamente per i dispositivi NMOS e PMOS studiati in questo lavoro di tesi. Confrontando il risultato ottenuto con i valori di $M_f(f)$ ottenuti in dispositivi appartenenti a precedenti nodi tecnologici (figura 3.33) non è possibile dedurre alcuna particolare tendenza nell'evoluzione di tale parametro con la riduzione della lunghezza minima di canale.

Conclusioni

Questo lavoro di tesi ha riguardato la caratterizzazione, sia in termini statici sia in termini di rumore, di dispositivi FinFET realizzati in una tecnologia con lunghezza minima di canale pari a 14 nm. Tale tecnologia presenta importanti innovazioni rispetto al processo convenzionale bulk CMOS e il suo studio, dal punto di vista statico e di rumore, è di fondamentale importanza da una parte per comprenderne i limiti, dall'altra per valutarne la capacità di rispettare le previsioni riguardanti l'evoluzione dei processi con lo scaling tecnologico. Dal punto di vita statico, oltre all'analisi delle curve corrente-tensione, particolare attenzione è stata rivolta all'analisi dei dati mediante i concetti di efficienza di transconduttanza e di coefficiente di inversione. A tal proposito, è stato effettuato uno studio del guadagno di tensione intrinseco A_{Vi} in funzione del coefficiente di inversione, che ha consentito di verificare che per lunghezze di canale minime, esso risulta confrontabile con i valori estratti da processi CMOS appartenenti a precedenti nodi tecnologici, secondo quanto previsto dalle regole di scaling. Parte della caratterizzazione statica è infine stata riservata all'analisi della tensione di soglia V_{TH} ed alla valutazione di diversi metodi presenti in letteratura per l'estrapolazione del suo valore.

Per la caratterizzazione in termini di rumore si è fatto riferimento a modelli che permettono la descrizione del contributo di rumore principale in un transistor ad effetto di campo con lunghezza di canale nanometrica, in particolare il rumore associato alla corrente di canale. Per quanto riguarda il contributo indipendente dalla frequenza (rumore termico di canale), il modello utilizzato tiene conto degli effetti di canale corto, trattati nel primo capitolo dell'elaborato, mediante il parametro α_w ed il coefficiente n_{sub} . Il rumore *flicker* è stato invece trattato facendo riferimento al modello proposto da McWorther. I dispositivi studiati sono caratterizzati da un'elevata componente di rumore 1/f (comunque confrontabile con quella già misurata in dispositivi appartenenti a precedenti nodi tecnologici CMOS), che ha reso dunque necessaria l'adozione di un opportuno setup di misura, caratterizzato da un amplificatore a transimpedenza con una larghezza di banda sufficientemente estesa da rendere possibile la misura del rumore termico di canale (almeno in alcune condizioni

di lavoro).

Mediante lo studio condotto è possibile affermare che la tecnologia FinFET da 14 nm, oggetto di questa tesi, non mostra un comportamento significativamente diverso rispetto ai processi bulk CMOS meno scalati e che essa è in linea con il processo di scaling tecnologico.

Esistono diverse attività future che possono essere sviluppate a partire da questo lavoro di tesi. Una, in modo particolare, riguarda lo studio della corrente di gate che, come già accennato nel capitolo 3 dell’elaborato, costituisce un aspetto particolarmente critico nei moderni processi microelettronici CMOS caratterizzati da spessori di ossido dell’ordine di pochi nanometri.

Un’altra attività interessante consiste nella valutazione delle prestazioni di rumore dei dispositivi analizzati in questa tesi in funzione del coefficiente di inversione del canale, al fine di permettere uno studio del comportamento del DUT nelle diverse regioni di debole, moderata e forte inversione. Infine, di particolare interesse risulta l’analisi di tecnologie CMOS SOI FD, che sono considerate un’alternativa ai FinFET per il proseguimento del processo di scaling.

Bibliografia

- [1] G. E. Moore, “Cramming More Components onto Integrated Circuits”, *Proc. IEEE*, **86** (1), Jan. 1998.
- [2] Semiconductor Industry Association (SIA), The National Technology Roadmap for Semiconductors. [Online.] Available WWW: <http://www.semtech.org/public/roadmap/index.htm>.
- [3] R. H. Dennard, F. H. Gaenslen, H. N. Yu, V. Leo Rideout, E. Bassous, A. R. LeBlanc, “Design of ion implanted MOSFETs with very small physical dimensions”, *IEEE Journal of Solid-State Circuits*, **9** (5), Oct. 1974.
- [4] R. S. Muller, T. I. Kamins, *Device electronics for integrated circuits*, 2nd ed. New York: John Wiley & Sons, 1986.
- [5] J. M. Rabaey, A. Chandrakasan, B. Nikolic, *Digital Integrated Circuits: A Design Perspective*, 2nd ed. Pearson Education, Inc. (Prentice Hall), Upper Saddle River, NJ, USA, 2003.
- [6] G. Baccarani, M. R. Wordeman, R. H. Dennard “Generalized scaling theory and its application to a 1/4 micrometer MOSFET design”, *IEEE Trans. Elect. Dev.*, **31** (4), 452-462, Apr. 2013.
- [7] S. H. Lo, D. A. Buchanan, Y. Taur, W. Wang, “Quantum-Mechanical Modeling of Electron Tunneling Current from the Inversion Layer of Ultra-Thin-Oxide nMOSFETs”, *IEEE Elect. Dev. Lett.*, **18** (5), May. 1997.
- [8] J. H. Stathis, “Reliability limits for the gate insulator in CMOS technology”, *IBM Journal of Research and Development*, **46** (2.3), 265-286, Mar. 2002.

- [9] G. He, Z. Sun, M. Liu, L. Zhang, *Scaling and limitation of Si-based CMOS High-k gate dielectrics for CMOS Technology*, 1st ed. Gang He and Z. Sun, 2012.
- [10] N. Z. Haron, “Why is CMOS scaling coming to an end?”, *Design and Test Workshop. IDT 2008. 3rd International*, 98-103, Dec. 2008.
- [11] J. G. Fossum, V. P. Trivedi, *Ultra thin body MOSFETs and FinFETs*, Cambridge University, United Kingdom, 2013.
- [12] L. D. Yau, “A simple theory to predict the threshold voltage of short channel IGFET’s”, *Solid State Electronics* **17**, 1059-1063, 1974.
- [13] I. L. Markov, “Limits on fundamental limits to computation”, EECS Department, University of Michigan, USA, 2014.
- [14] K. Kuhn, C. Kenyon, A. Kornfeld, M. Liu, A. Maheshwari, W. K. Shih, S. Sivakumar, G. Taylor, P. VanDerVoorn, K. Zawadzki, “Managing process variation in Intel’s 45 nm CMOS technology”, *Intel Technology Journal Comput. Electron.*, **12**, (2), 93-109, 2008.
- [15] R. C. Jaeger, *Introduction to Microelectronic Fabrication*, Volume 5, 2nd ed. Prentice Hall, New Jersey, USA, 2002.
- [16] J.L Hoyt et al., “Strained silicon MOSFET Technology”, *Electron Dev. Meeting*, 23-26, Dec. 2002.
- [17] M. J. Kumar, M. Siva, “The Ground Plane in buried oxide for controlling Short-Channel Effects in nanoscale SOI MOSFETs”, *IEEE Trans. Elect. Dev.*, **55** (6), 1554-1557, Jun. 2008.
- [18] M. Fujiwara, T. Morooka, N. Yasutake, K. Ohuchi, “Impact of BOX scaling on 30 nm gate length FD SOI MOSFET”, *2005 IEEE International SOI Conference* , 180-182, Oct. 2005.
- [19] C. Chuang et al., “Scaling planar silicon devices”, *IEEE Circuits Devices Mag.*, 6-19, Feb. 2004.
- [20] E. J. Nowak, I. Aller, T. Ludwig, K. Keunwoo, “Turning silicon on its edge”, *IEEE Circ. Dev. Mag.* **20** (1), 20-31, Jan. 2004.
- [21] L. Goeppert, “The Amazing Vanishing Transistor Act”, *IEEE Spec.*, 28-33, Oct. 2002.

- [22] T. J. King Liu, "FinFET History, Fundamentals and Future", 2012 Symposium on VLSI Technology Short Course, June 2012.
- [23] Y. P. Tsividis, *Operation and Modeling of the MOS Transistor*, 2nd ed. New York, NY, USA: Mc Graw-Hill, 1999.
- [24] G. Anelli, "Analog Design in ULSI CMOS Processes", *Proceedings 2004 10th Workshop on Electronics for LHC Experiments and Future Experiments*, Boston, Sep. 2004.
- [25] S. C. Terry et al., "Comparison of BSIM3v3 and EKV MOSFET model for a 0.5 μm CMOS process and implications for analog circuit design", *IEEE Trans. Nucl. Sci.*, **50** (4), 915-920, Aug. 2003.
- [26] T.K. Liu et al., "MuGFET Carrier Mobility and Velocity: Impacts of Fin Aspect Ratio, Orientation and Stress", *IEEE Elec. Dev. Meeting*, Dec. 2010.
- [27] C. C. Enz and E. A. Vittoz, *Charge-Based MOS transistor modeling: The EKV model for low-power and RF IC design*. Wiley, New York, NY, 2006.
- [28] F. Maloberti, *Analog Design for CMOS VLSI System*. Kluwer Academic Publishers, New York, USA, 2001.
- [29] M. Garg et al., "Scaling impact on analog performance of sub-100 nm MOSFETs for mixed mode applications", in Proc. *33rd Eur. Solid-State Device Res. Conf.*, 2003, 371-374.
- [30] A. Ghetti, E. Sangiorgi, J. Bude, T.W. Sorsch, G. Weber, "Tunneling into Interface States as Reliability Monitor for Ultrathin Oxides", *IEEE Trans. Electron Devices*, **47** (12), 2358-2365, Dec. 2000
- [31] T. Rudenko et al., "On the MOSFET Threshold Voltage Extraction by Transconductance and Transconductance-to-Current Ratio Change Methods: Part I'Effect of Gate-Voltage-Dependent Mobility". *IEEE Trans. Electr. Dev.*, **58** (12), 4172-4179, Dec. 2011.
- [32] A. B. Fowler and A. M. Hartstein, "Techniques for determining threshold", *Surf. Sci.*, **98** (1)-(3), 169-172, Aug. 1980.
- [33] G. Ghibaudo, "New method for the extraction of MOSFET parameters," *Electron. Lett.*, **24** (9), 543-545, Apr. 1988.

- [34] D. Flandre, V. Kilchytska, T. Rudenko, "gm/Id Method for Threshold Voltage Extraction Applicable in Advanced MOSFETs With Nonlinear Behavior Above Threshold", *IEEE Elect. Dev. Lett.*, **31** (9), 930-932, Sep. 2010.
- [35] K. Aoyama, "A method for extracting the threshold voltage of MOSFETs based current components", *Simulation of Semiconductor Devices and Processes*, **6**, 118-121, Sep. 1995.
- [36] T. Rudenko, D. Flandre, V. Kilchytska, A. Nazarov, "Influence of Drain Voltage on MOSFET Threshold Voltage Determination by Transconductance Change and gm/Id Methods", *Conference on Ultimate Integration on Silicon (ULIS), 2011 12th International*, 1-4, 2011.
- [37] M. Manghi, "Gate Current Noise in Ultrathin Oxide MOSFETs and its Impact of the Performance of Analog Front-End Circuit", *IEEE Trans. Nucl. Sci.*, **55** (4), 2399-2407, Aug. 2008.
- [38] M. Manghisoni, L. Gaioni, L. Ratti, V. Re, G. Traversi, "Assessment of a Low-Power 65 nm CMOS Technology for Analog Front-End Design", *IEEE Tran. Nuc. Sci* **61**, (1) Feb. 2014.
- [39] A. L. McWhorter, "1/f noise and germanium surface properties", in *Semiconductor Surface Physics*. Philadelphia, PA, USA: Univ. of Pennsylvania Press, 1957.
- [40] F. N. Hooge, "1/f noise," *Physica*, **83B**, 14, 1976.
- [41] K. K. Hung et al., "A unified model for the flicker noise in metal-oxide-semiconductor field-effect transistors", *IEEE Trans. Elect. Dev.* **37** (3), 654-665, Aug. 2002.