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

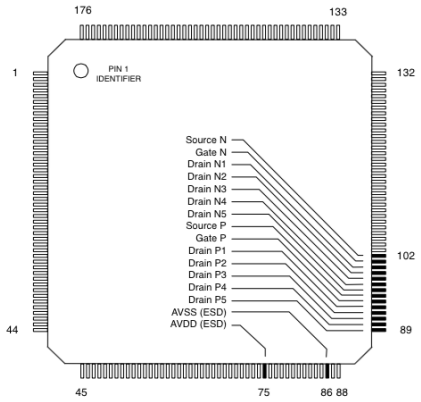
I dispositivi studiati sono FinFET a canale N e a canale P, caratterizzati da differenti dimensioni di gate. I DUT sono stati resi disponibili sia come *naked dice*, successivamente connessi tramite *wire bonding* ad un pakage LCC44, sia direttamente nel pakage TQFP 176L. Per rendere possibile la caratterizzazione 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 μm. 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 $8 \cdot 74$ 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 $V_{DD} = 0.8$ V.

NMOS				
W [μ**m**]	**L** [**nm**]	$\mathbf{n_d}$	**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 [μ**m**]	**L** [**nm**]	$\mathbf{n_d}$	**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.



Pin #	Pin Name	Description
North side		
102	SOURCEN	Common source of NMOS
101	GATEN	Common gate of NMOS
100	DN1	Drain N1
99	DN2	Drain N2
98	DN3	Drain N3
97	DN4	Drain N4
96	DN5	Drain N5
95	SOURCEP	Common source of PMOS
94	GATEP	Common gate of PMOS
93	DP1	Drain P1
92	DP2	Drain P2
91	DP3	Drain P3
90	DP4	Drain P4
89	DP5	Drain P5
East side		
86	AVSS	Analog ground (ESD)
75	AVDD0p8	Analog power (ESD)

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

Pin Name	Description
GATEP	Common gate of PMOS (shorted to pad ring one)
DP9	Drain P9
DP10	Drain P10
DP11	Drain P11
DP12	Drain P12
SOURCEP	Common source of PMOS (shorted to pad ring one)
DP8	Drain P8
DP7	Drain P7
DP6	Drain P6
DN8	Drain N8
DN7	Drain N7
DN6	Drain N6
GATEN	Common gate of NMOS (shorted to pad ring one)
DN9	Drain N9
DN10	Drain N10
DN11	Drain N11
DN12	Drain N12
SOURCEN	Common source of NMOS (shorted to pad ring one)

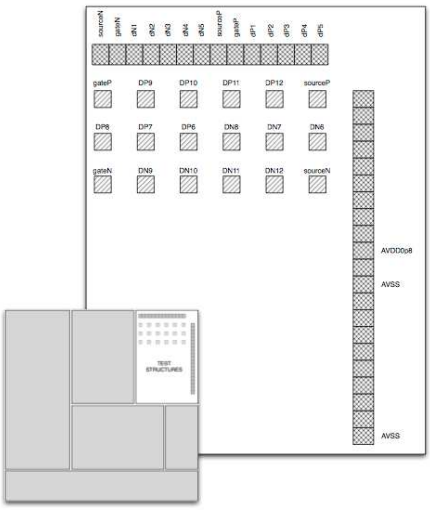


Figura 2.2: bonding pads dei chip connessi al pakage LCC44.

[1] I *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-Measurament 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 Analayzer*. 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.

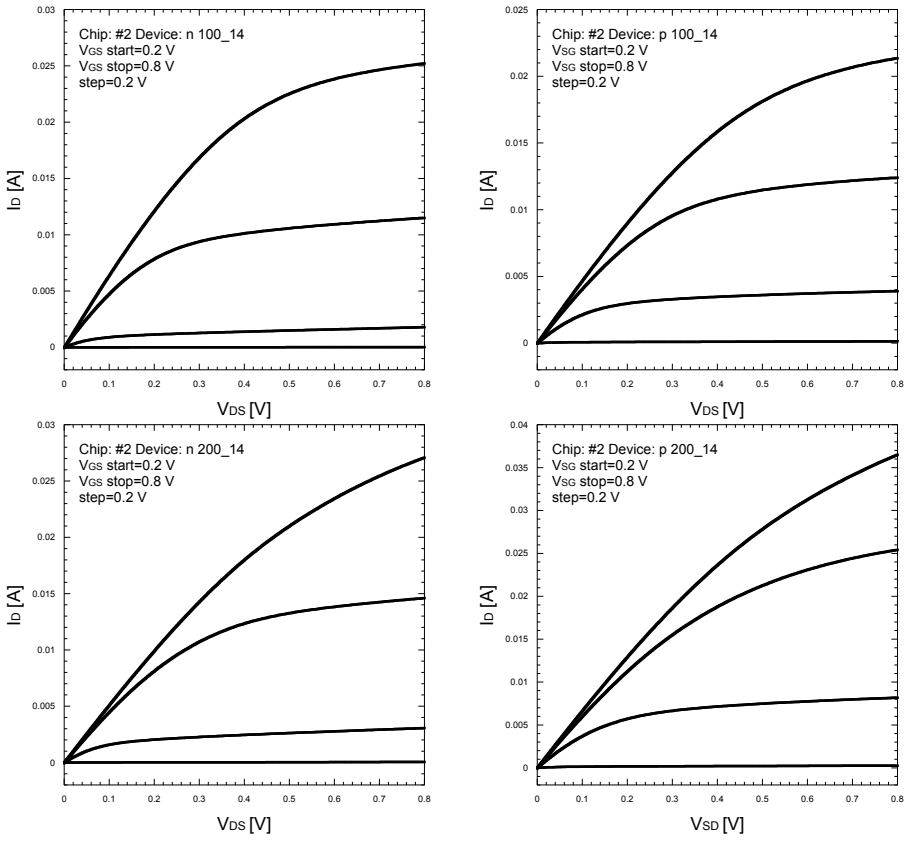


Figura 2.3: corrente di drain I_D in funzione della tensione drain-source V_{DS} con V_{GS} come parametro per dispositivi a canale N (a sinistra) e a canale P (a destra) con lunghezza di canale $L = 14$ nm.

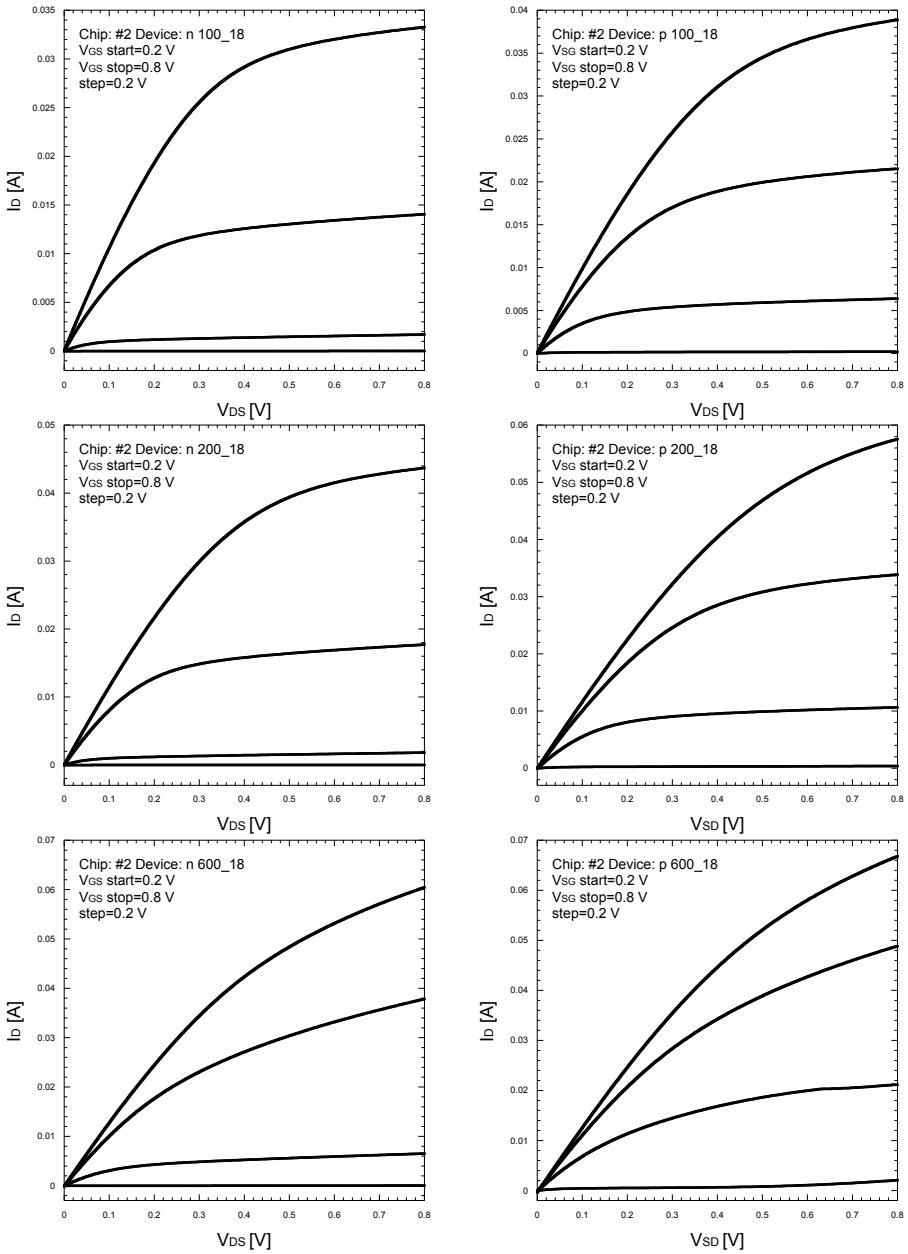


Figura 2.4: corrente di drain I_D in funzione della tensione drain-source V_{DS} con V_{GS} come parametro per dispositivi a canale N (a sinistra) e a canale P (a destra) con lunghezza di canale $L = 18$ nm.

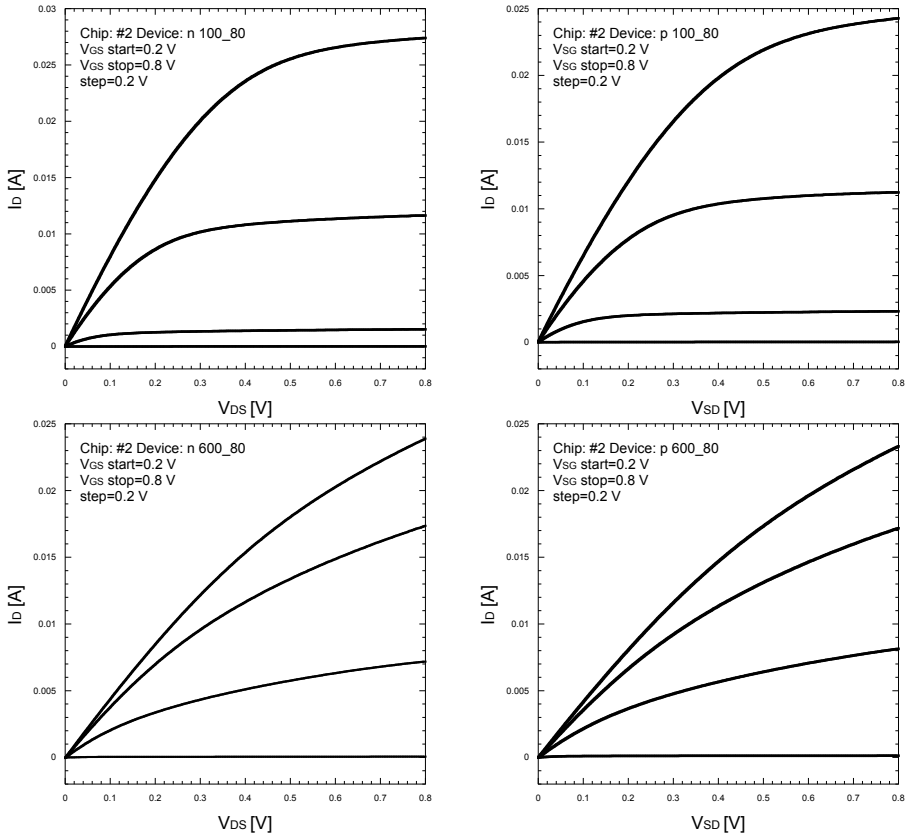


Figura 2.5: corrente di drain I_D in funzione della tensione drain-source V_{DS} con V_{GS} come parametro per dispositivi a canale N (a sinistra) e a canale P (a destra) con lunghezza di canale $L = 80$ nm.

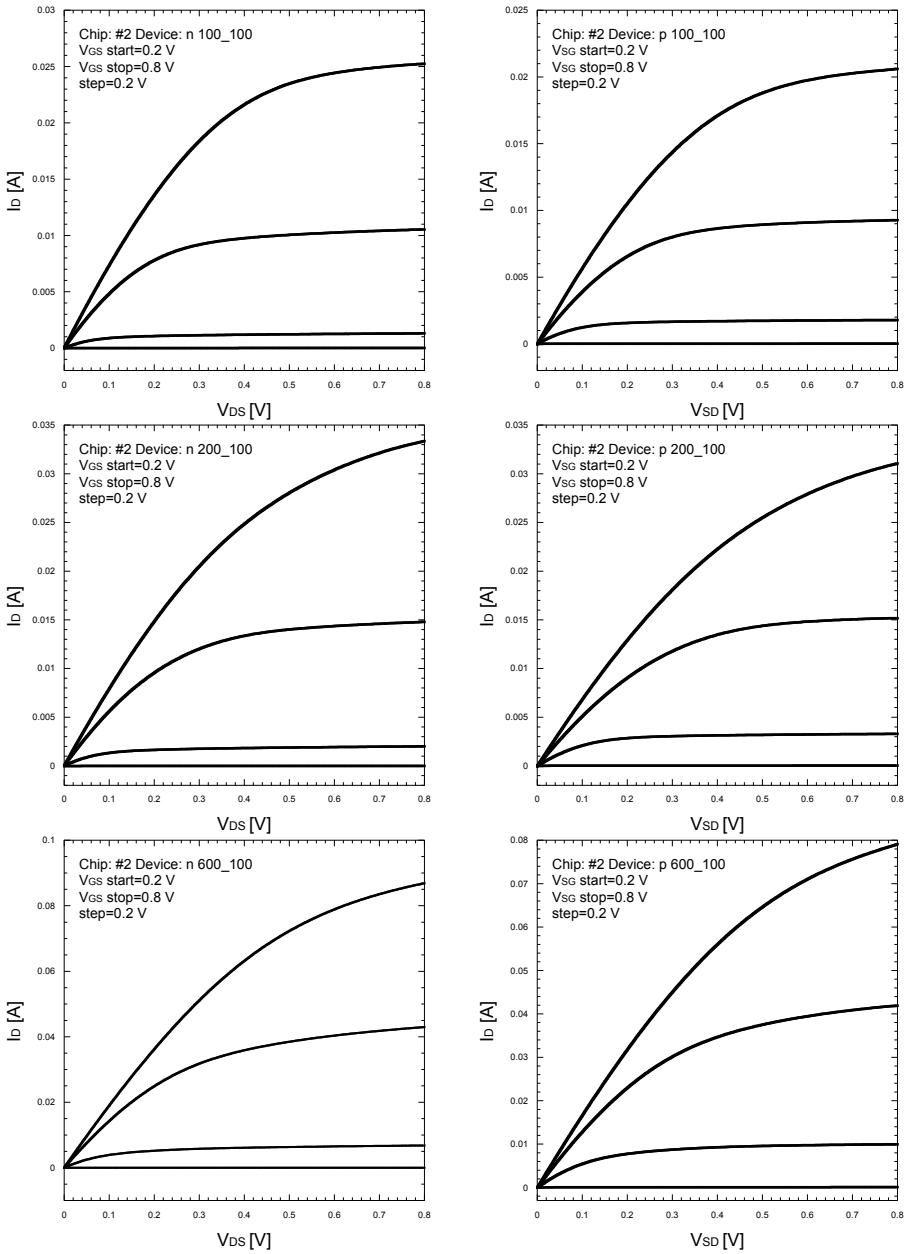


Figura 2.6: corrente di drain I_D in funzione della tensione drain-source V_{DS} con V_{GS} come parametro per dispositivi a canale N (a sinistra) e a canale P (a destra) con lunghezza di canale $L = 100$ nm.

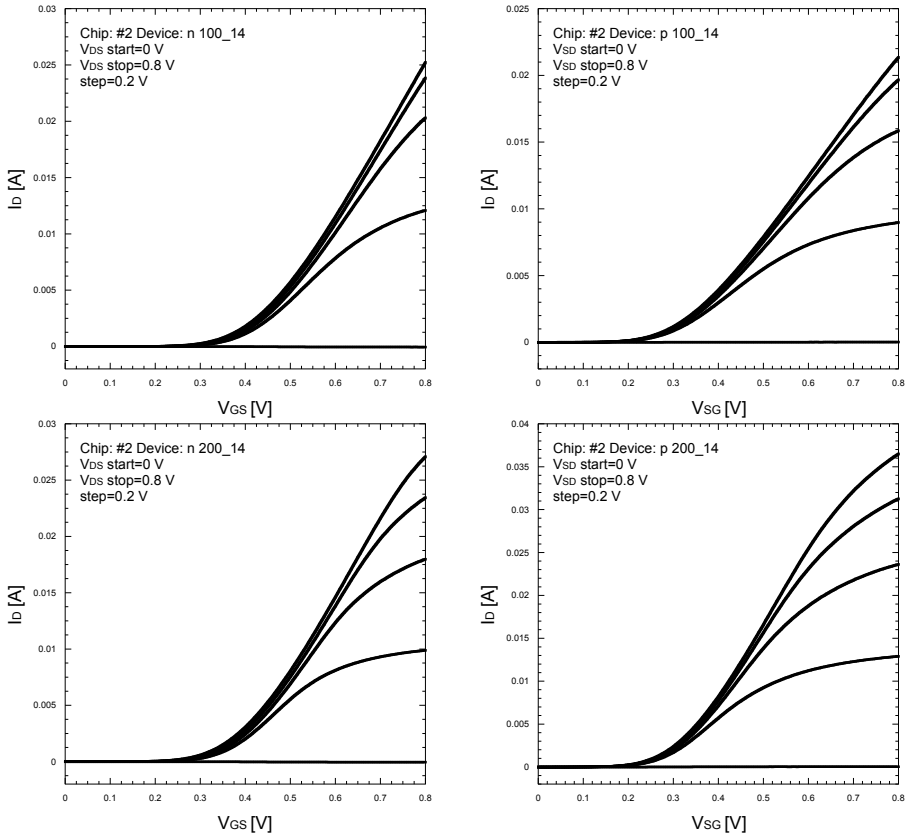


Figura 2.7: corrente di drain I_D in funzione della tensione gate-source V_{GS} con V_{DS} come parametro per dispositivi a canale N (a sinistra) e a canale P (a destra) con lunghezza di canale $L = 14$ nm.

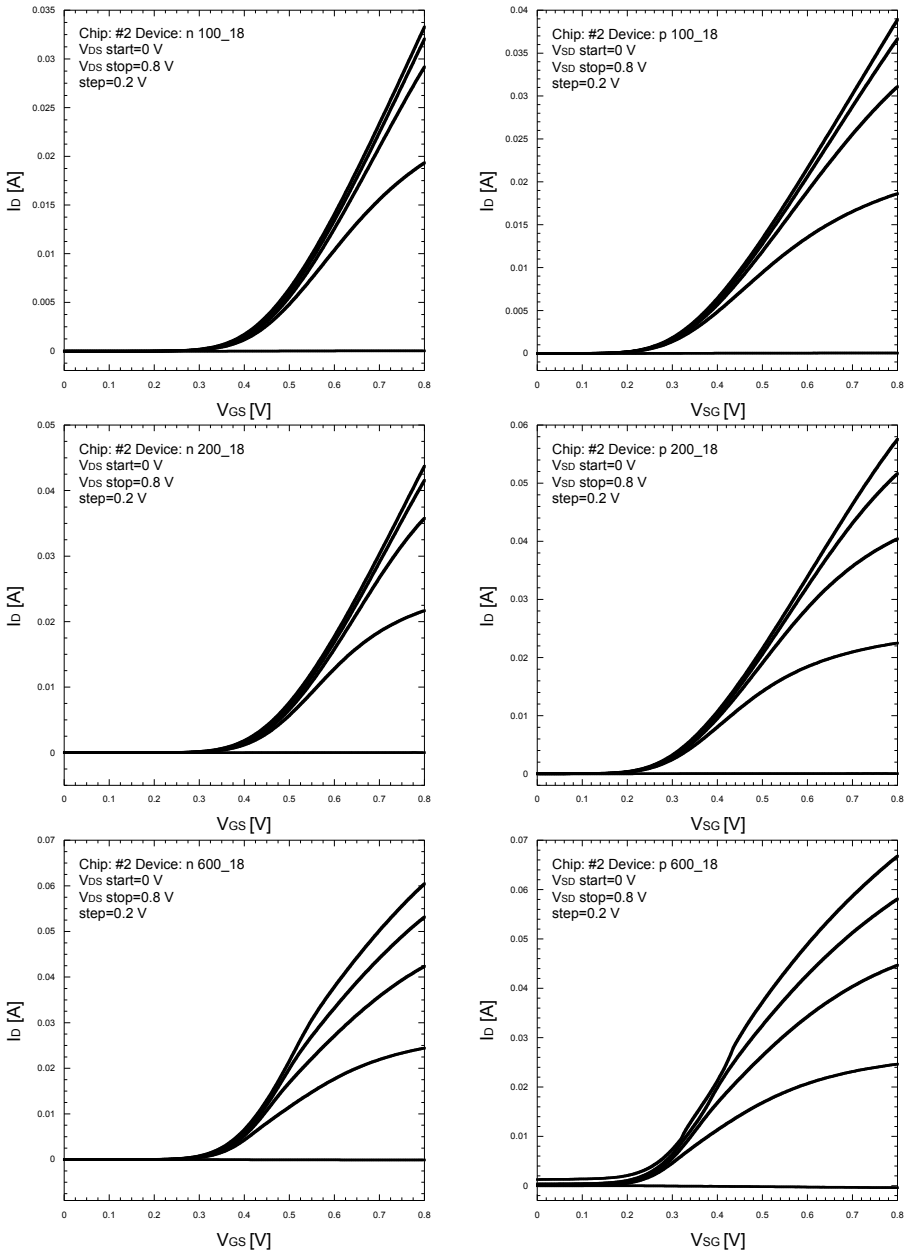


Figura 2.8: corrente di drain I_D in funzione della tensione gate-source V_{GS} con V_{DS} come parametro per dispositivi a canale N (a sinistra) e a canale P (a destra) con lunghezza di canale $L = 18$ nm.

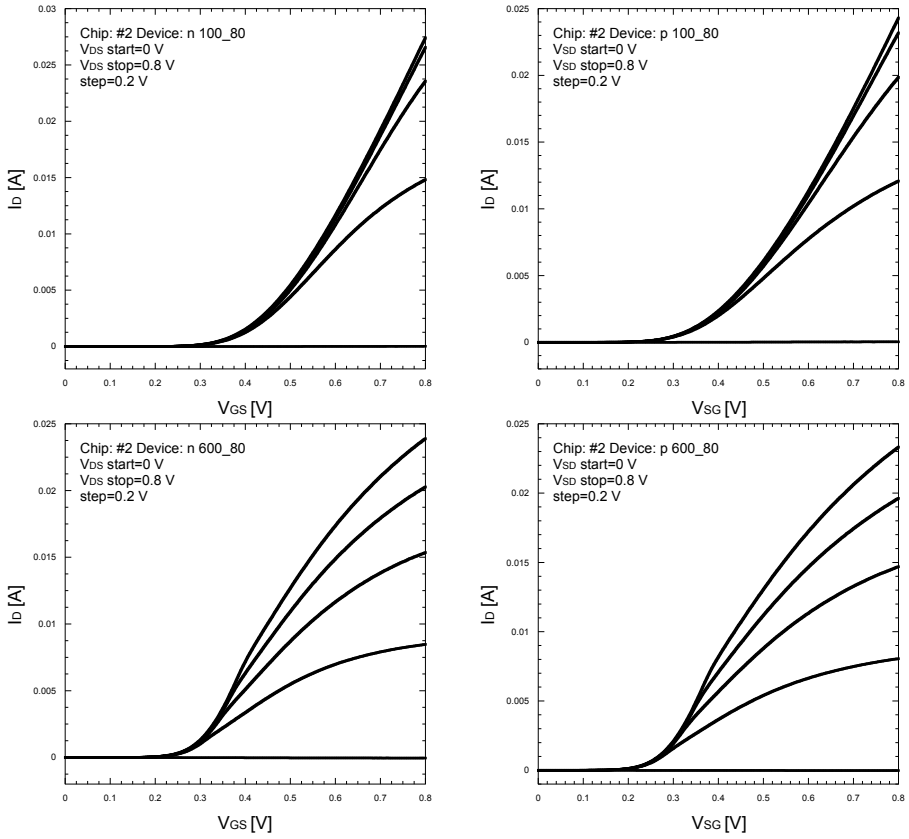


Figura 2.9: corrente di drain I_D in funzione della tensione gate-source V_{GS} con V_{DS} come parametro per dispositivi a canale N (a sinistra) e a canale P (a destra) con lunghezza di canale $L = 80$ nm.

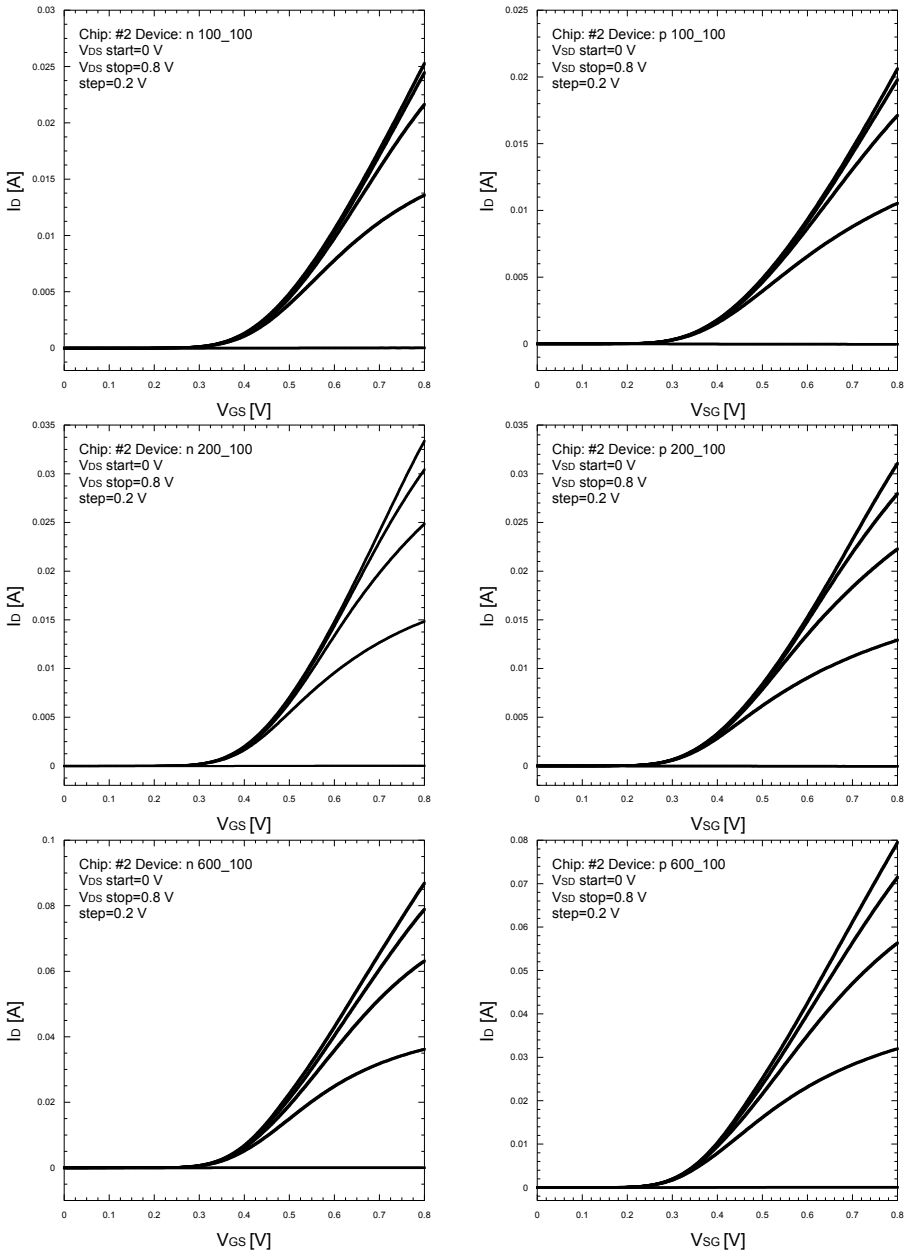


Figura 2.10: corrente di drain I_D in funzione della tensione gate-source V_{GS} con V_{DS} come parametro per dispositivi a canale N (a sinistra) e a canale P (a destra) con lunghezza di canale $L = 100$ nm.

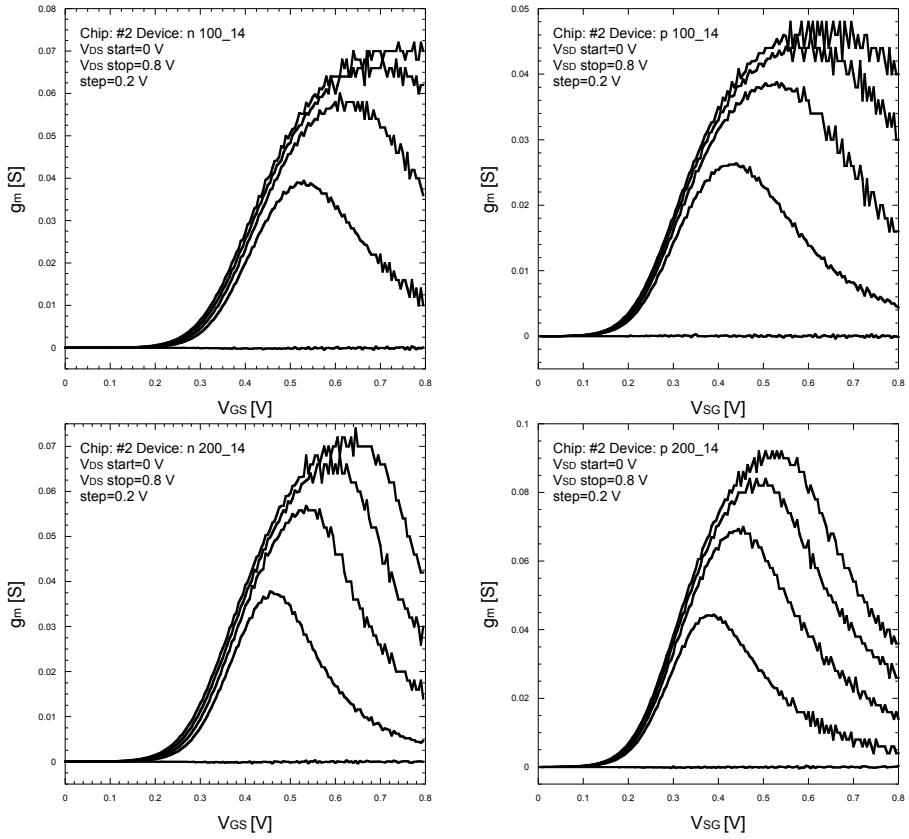


Figura 2.11: transconduttanza g_m in funzione della tensione V_{GS} con V_{DS} come parametro per dispositivi a canale N (a sinistra) e a canale P (a destra) con lunghezza di canale $L = 14$ nm.

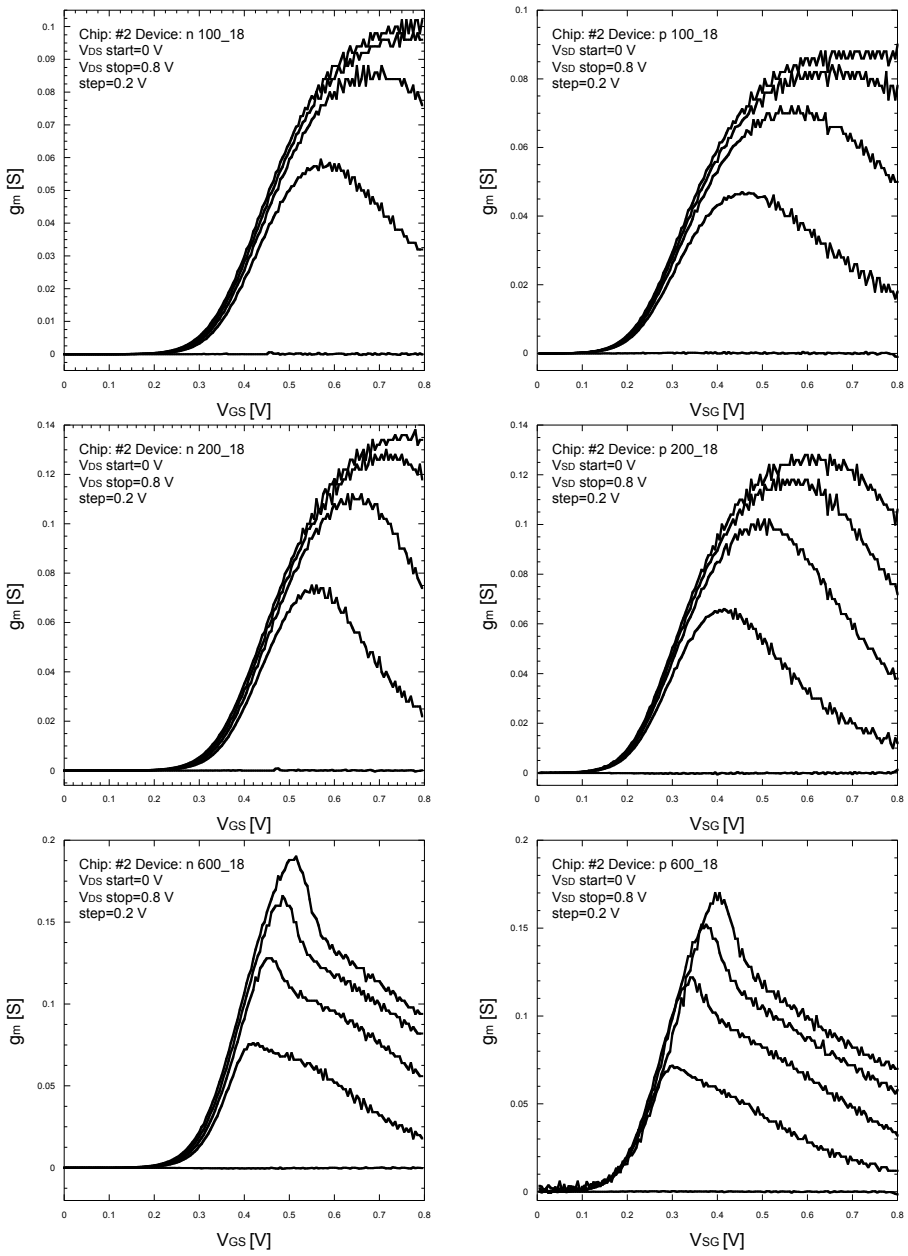


Figura 2.12: transconduttanza g_m in funzione della tensione V_{GS} con V_{DS} come parametro per dispositivi a canale N (a sinistra) e a canale P (a destra) con lunghezza di canale $L = 18$ nm.

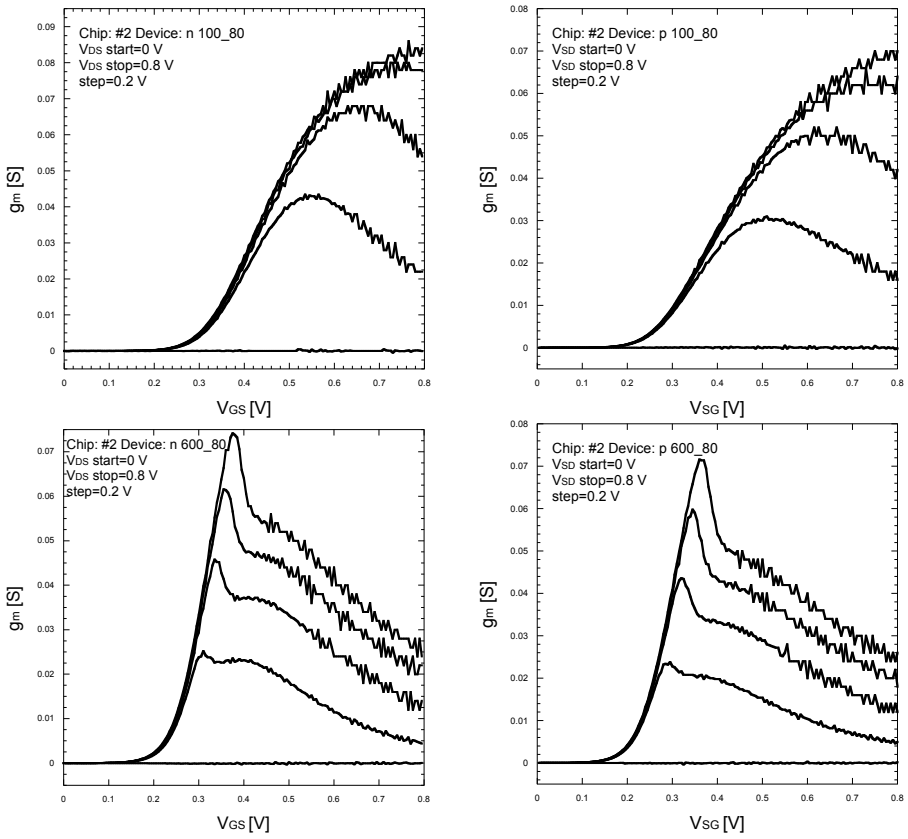


Figura 2.13: transconduttanza g_m in funzione della tensione V_{GS} con V_{DS} come parametro per dispositivi a canale N (a sinistra) e a canale P (a destra) con lunghezza di canale $L = 80$ nm.

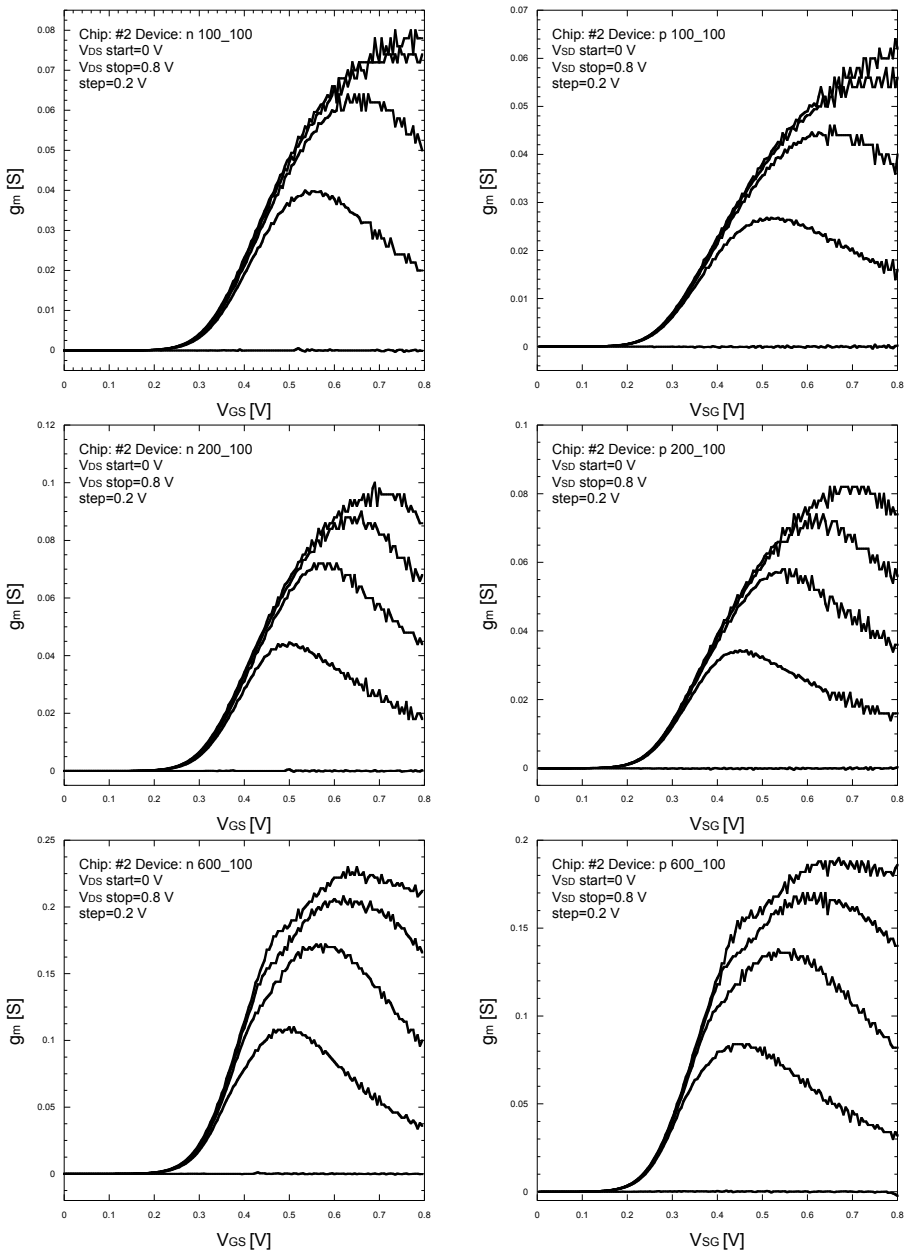


Figura 2.14: transconduttanza g_m in funzione della tensione V_{GS} con V_{DS} come parametro per dispositivi a canale N (a sinistra) e a canale P (a destra) con lunghezza di canale $L = 100$ nm.

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:

$$g_m = \frac{I_D}{n_{sub} V_t}, \tag{2.1}$$

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:

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

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

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

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

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

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]:

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

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

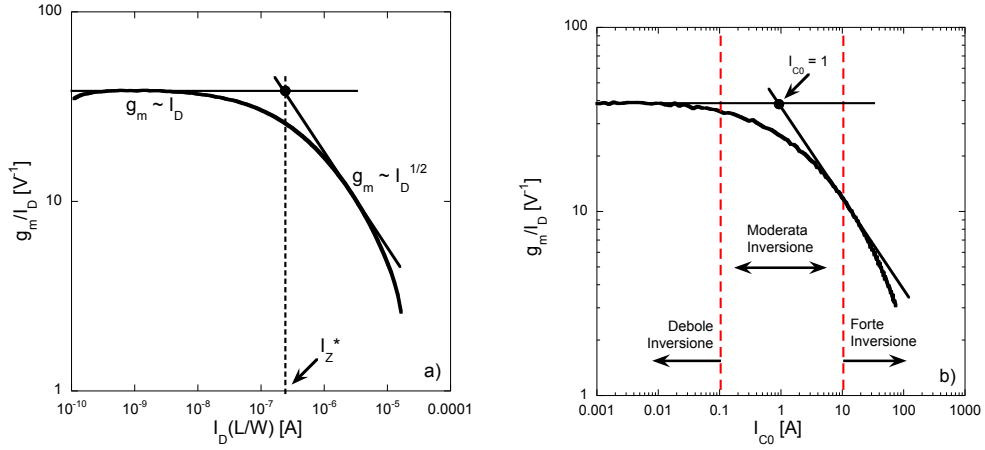


Figura 2.15: efficienza di transconduttanza, ricavata da misure sui FinFET, in funzione della a) corrente normalizzata di drain e b) del coefficiente di inversione I_{C0}.

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]:

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

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 colloca-

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. ll coefficiente di inversione, I_{C0}, viene definito come [25]:

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

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:

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

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.

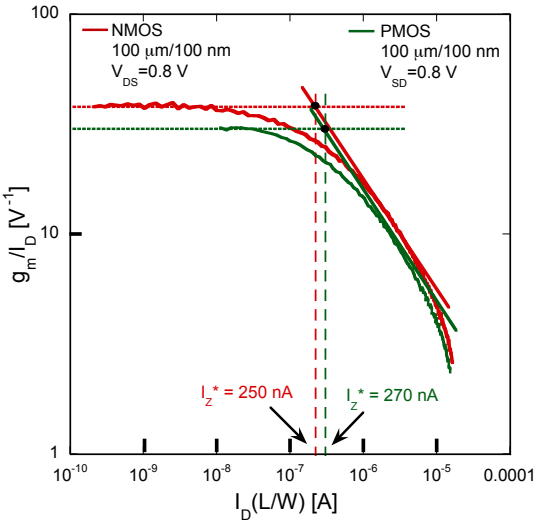


Figura 2.16: estrazione della corrente caratteristica di drain normalizzata I_Z^* per un FinFET a canale N ed un FinFET a canale P con uguali dimensioni di gate.

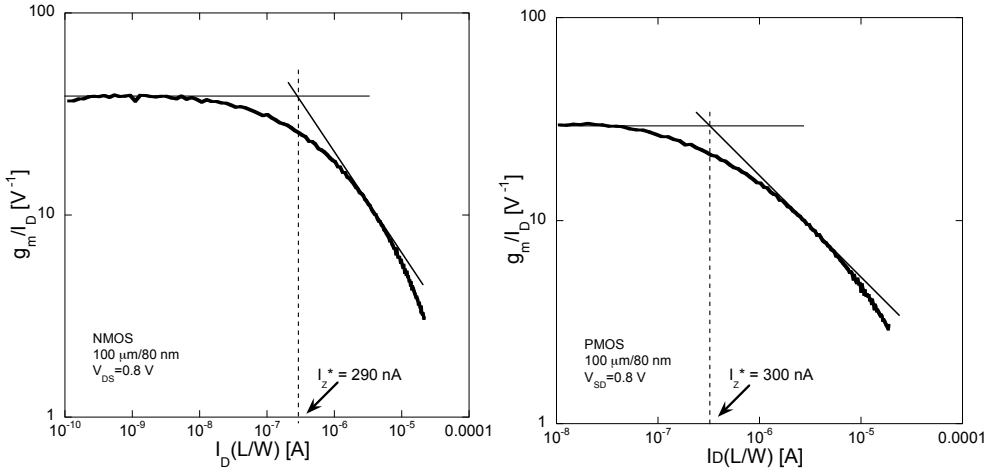


Figura 2.17: estrazione della corrente caratteristica I_Z^* per un FinFET a canale N (a sinistra) e per uno a canale P (a destra) con $W/L = 100\mu m/80nm$.

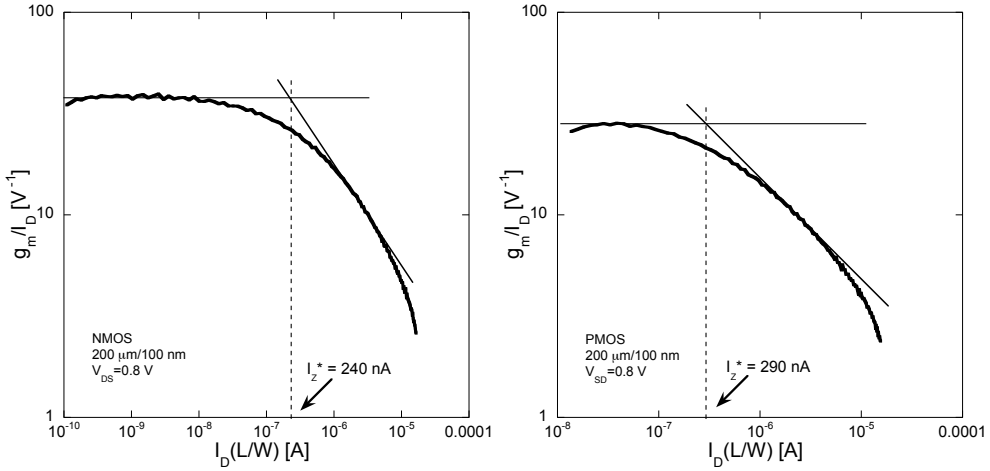


Figura 2.18: estrazione della corrente caratteristica I_Z^* per un FinFET a canale N (a sinistra) e per uno a canale P (a destra) con $W/L = 200\mu m/100nm$.

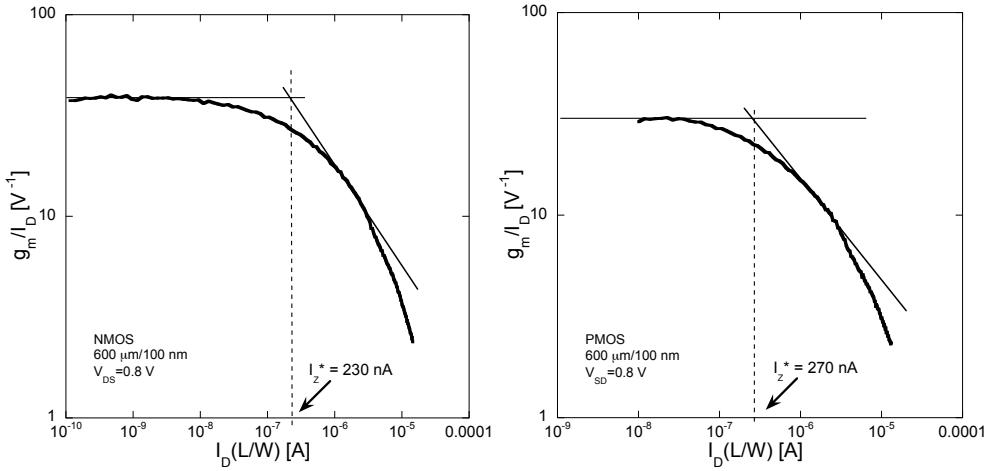


Figura 2.19: estrazione della corrente caratteristica I_Z^* per un FinFET a canale N (a sinistra) e per uno a canale P (a destra) con $W/L = 600\mu m/100nm$.

	NMOS		**PMOS**	
	n_{sub}	I_Z^* [nA]	n_{sub}	I_Z^* [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 dispostivi 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}:

$$A_{Vi} = \frac{g_m}{g_{ds}}. \tag{2.9}$$

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]:

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

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} constanti, 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]:

$$g_m = \frac{I_D}{n_{sub} \frac{K_B T}{q}}; \tag{2.11}$$

$$g_{ds} = \lambda_{wi} I_D, \tag{2.12}$$

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:

$$g_{ds} = \lambda I_D, \tag{2.13}$$

dove λ può essere approssimata come:

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

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

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

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-

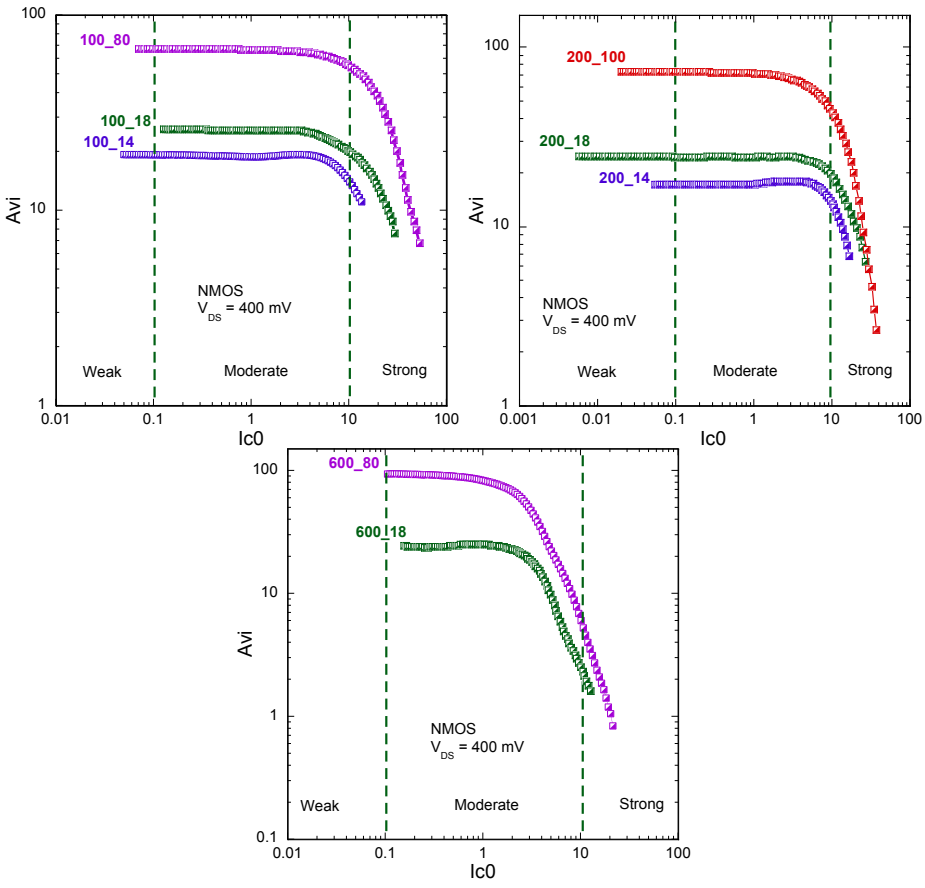


Figura 2.20: guadagno intrinseco A_{Vi} in funzione della lunghezza di gate L_G per dispositivi NMOS appartenenti a diversi nodi tecnologici e polarizzati a $I_{C0} = 0.1$.

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} constante con lo scaling è considerato una delle principali sfide nel progetto delle tecnologie nanometriche [29]. È possibile infatti dimostrare che [24]:

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

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 elettrostatica 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.

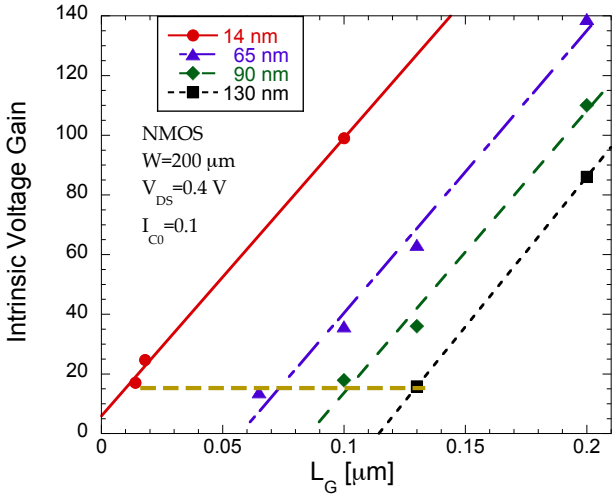


Figura 2.21: guadagno intrinseco A_{Vi} in funzione della lunghezza di gate L_G per dispositivi NMOS appartenenti a diversi nodi tecnologici e polarizzati nella regione di inversione $I_{C0} = 0.1$.

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'estrapolazione 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 apporossimativamente 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} ($\mathrm{d}^2(\log I_{DS})/\mathrm{d}V_{GS}^2$) 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:

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

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 bipolare[2]. 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

[2]Ipotizzando 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.

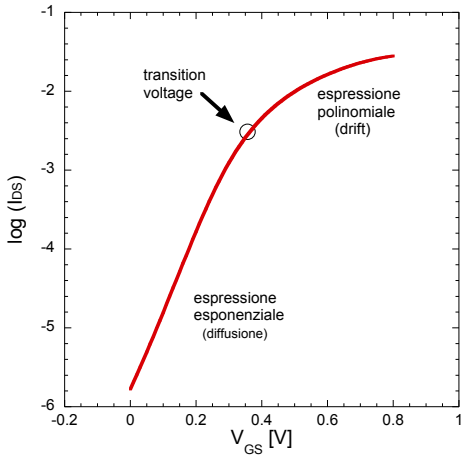


Figura 2.22: $\log(I_D)$ in funzione di V_{GS}.

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

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

in saturazione da:

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

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, $\mathrm{d}(\log I_{DS})/\mathrm{d}V_{GS}$ (figura 2.23(a));

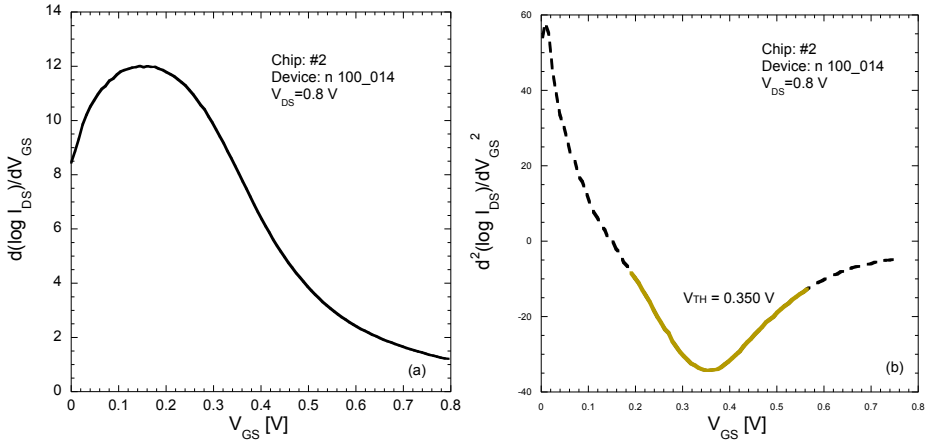


Figura 2.23: rappresentazione grafica del metodo SDLM per un FinFET con W/L=100 μm/14 nm .

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

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 $\mathrm{d}(\log I_{DS})/\mathrm{d}V_{GS}$ corrisponde al rapporto g_m/I_{DS}:

$$\frac{g_m}{I_{DS}} = \frac{1}{I_{DS}}\frac{\mathrm{d}I_{DS}}{\mathrm{d}V_{GS}} = \frac{\mathrm{d}(\log I_{DS})}{\mathrm{d}V_{GS}}. \tag{2.20}$$

Calcolando la derivata della (2.20) si ottiene:

$$\frac{\mathrm{d}(g_m/I_{DS})}{\mathrm{d}V_{GS}} = \frac{\mathrm{d}^2(\log I_{DS})}{\mathrm{d}V_{GS}^2}. \tag{2.21}$$

In definitiva la tensione di soglia corrisponde al valore massimo di $-\frac{\mathrm{d}(g_m/I_{DS})}{\mathrm{d}V_{GS}}$. 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 $\mathrm{d}g_m/\mathrm{d}V_{GS}$ all'aumentare di V_{DS}, corrisponda

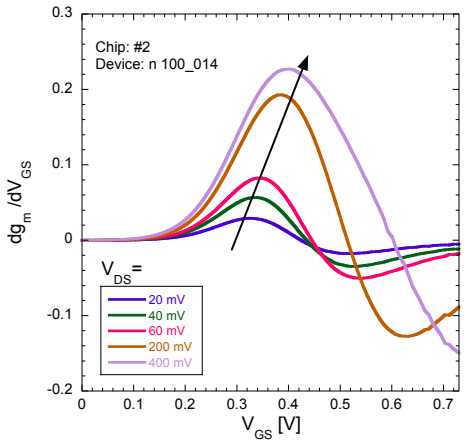


Figura 2.24: $\mathrm{d}g_m/\mathrm{d}V_{GS}$ in funzione della tensione di gate V_{GS} per differenti valori della tensione di drain V_{DS}.

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 $\mathrm{d}g_m/\mathrm{d}V_{GS}$ deve essere applicato a basse tensioni di drain mentre quello della derivata seconda $(\mathrm{d}(\log I_{DS})/\mathrm{d}V_{GS})$ 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.

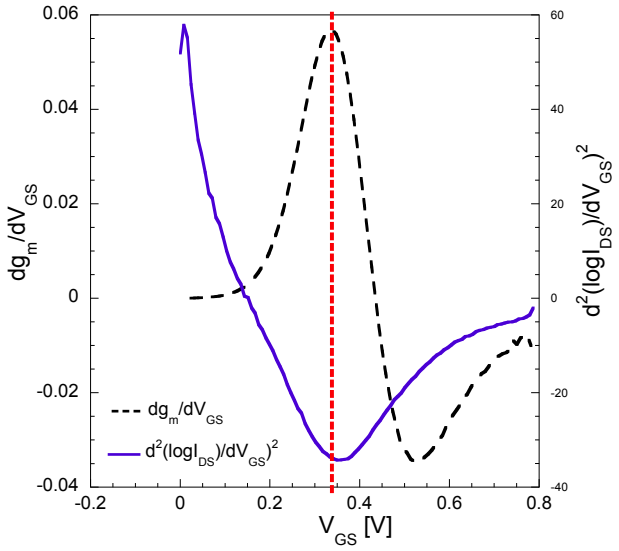


Figura 2.25: confronto tra due diversi metodi di estrazione del valore della tensione di soglia.

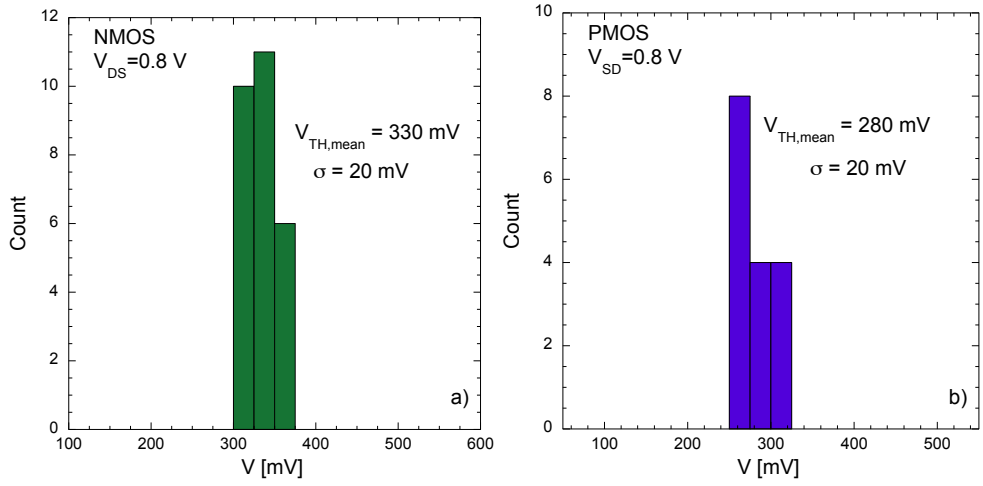


Figura 2.26: distribuzione dei valori estratti per la tensione di soglia per i dispositivi a) NMOS e b) PMOS.

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 propietà 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:

$$S = [\mathrm{d}\log(I_{DS})/\mathrm{d}V_{GS}]^{-1} = \left(\frac{n_{sub}k_BT}{q}\right)\ln 10. \tag{2.22}$$

In un transistor ideale $n_{sub} = 1$ e $S = (K_BT/q)\ln 10 \simeq 60$ 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 figure 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

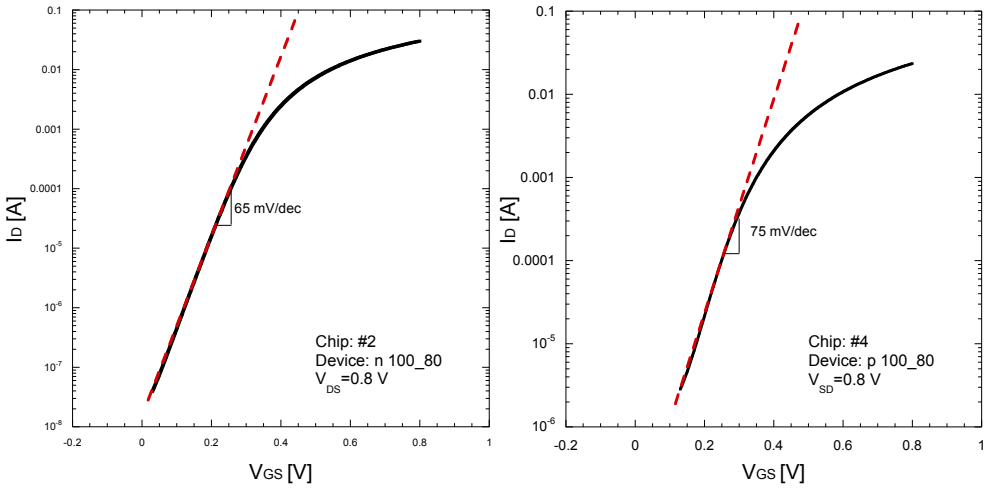


Figura 2.27: estrazione della pendenza di sottosoglia S.

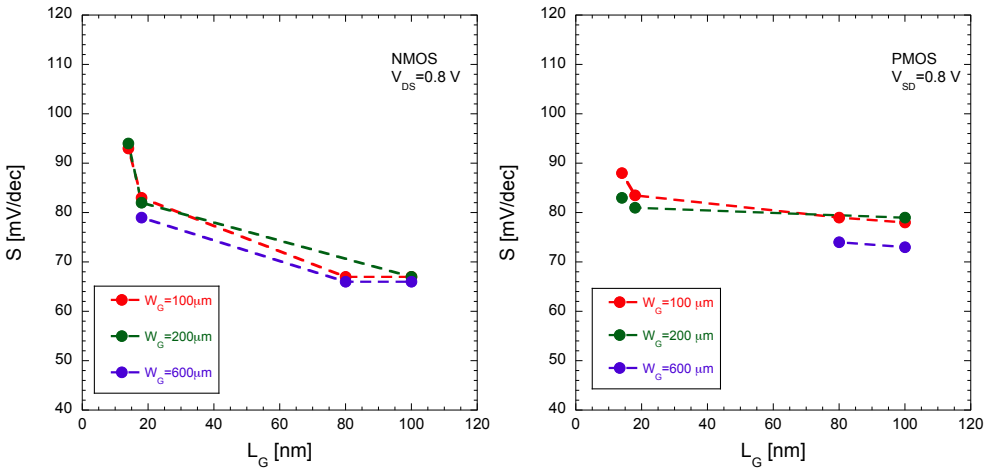


Figura 2.28: dipendenza di S dalla la lunghezza di canale L_G.

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:

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

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 conseguente-

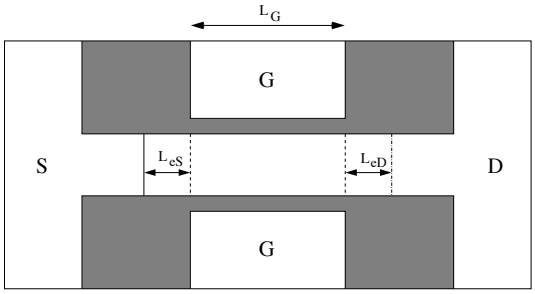


Figura 2.29: struttura schematizzata di un DG-MOSFET che indica le porzioni non drogate L_{eS} e L_{eD}.

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

$$L_{eff} \approx L_G. \tag{2.24}$$

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 appartamenti ad un nodo tecnologico precedente. Si nota che a parità di lun-

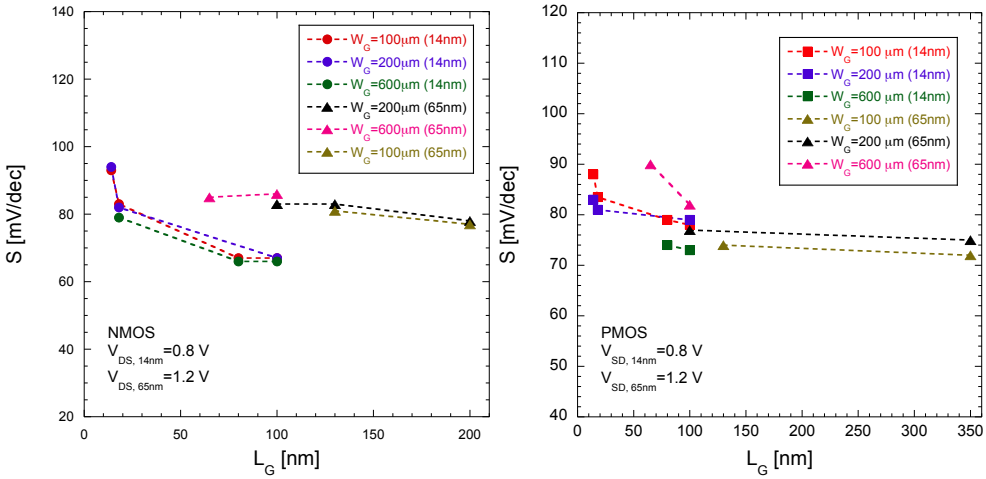


Figura 2.30: confronto della pendenza di sottosoglia S tra i dispositivi realizzati in tecnologia da 14 nm e quelli appartenenti ad una tecnologia da 65 nm.

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.