Università degli Studi di Padova

Facoltà di Ingegneria

Corso di Laurea in Ingegneria Dell'Informazione

Tesi di Laurea Triennale

Dissipazione di Potenza nei Circuiti CMOS: Origini e Tecniche per la Riduzione

Relatore: Alessandro Paccagnella

Laureando: Giovanni Bruni

27/09/2011

Questo documento è stato scritto in LaTeX su Debian GNU/Linux.

Indice

1. Introduzione

Il consumo di potenza da parte dei dispositivi elettronici sta diventando col tempo un fattore sempre più importante: il massiccio aumento dell'utilizzo di dispositivi di tipo *mobile*, come *smartphone*, *netbook* e gli "ultimi" *tablet* costringe i produttori di *hardware* a diminuire le richieste in termini di potenza dei loro prodotti, per poter avere una maggiore autonomia delle batterie. L'elettronica infatti non si limita più solo al mondo dei computer, ma ha conquistato altri settori: dal *Wi-Fi* al *GPS* degli *smartphone*, dai sensori delle fotocamere digitali ai *touchscreen* dei *tablet*, dalla strumentazione medica ai *MEMS*, cioè quei microsistemi in cui si integrano componenti miniaturizzati di varia natura ingegneristica.

Per poter ottenere una sempre maggiore miniaturizzazione di tali sistemi (si pensi alla grande quantità di elementi presenti in un singolo *smartphone*) e, parallelamente, un maggior numero di funzionalità di un singolo *chip*, si tende a diminuire sempre più le dimensioni dei dispositivi elementari per poterne aumentare il numero (*scaling*).

Riguardo lo *scaling* dei MOSFET, questo ha sempre seguito la nota *legge di Moore*, di cui se ne riporta una simpatica rappresentazione in Figura 1.1 [1]:

Ogni due anni il numero di transistor di un processore raddoppia.

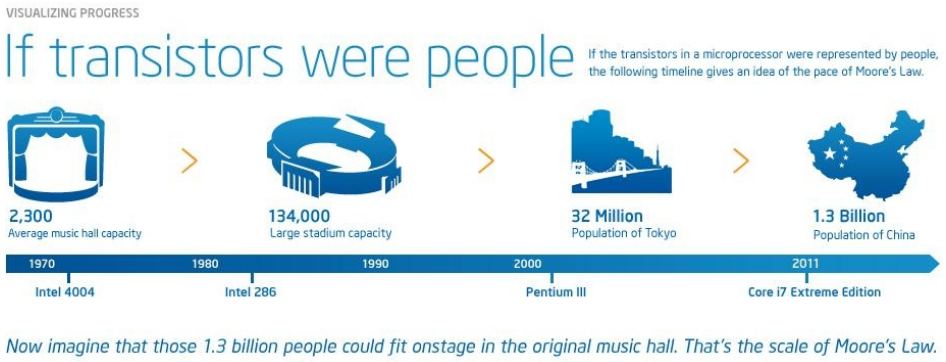


Figura 1.1: Legge di Moore

La riduzione delle dimensioni dei *transistor* tuttavia è sottoposta a vincoli di varia natura:

Progettuale - a causa ad esempio dell'aumento della complessità delle interconnessioni;

Termica - poiché un aumento del numero di transistor operanti nello stesso *chip* porta ad un aumento anche del calore dissipato e della temperatura del *chip*;

Robustezza - dato che, ad esempio, aumentano gli accoppiamenti fra le varie componenti del circuito e quindi aumentano i disturbi.

L'aumento del numero di transistor in un singolo circuito integrato, ovviamente, non significa solamente una maggiore capacità computazionale, ma, come è stato detto precedentemente, soprattutto una crescita costante di ciò che un singolo *chip* può offrire, come ad esempio accade nei processori *multi-core*.

Tuttavia un aumento del numero di transistor in un unico dispositivo ha come principale effetto negativo l'aumento del consumo di potenza, che risulta dannoso soprattutto nei sistemi portatili alimentati a batterie. Per mantenere basso il consumo di questi apparati si è proceduto all'inizio a diminuire di circa il 30% la tensione di alimentazione (V_{DD}) per ogni nuova generazione: questo tuttavia ha costretto a diminuire la tensione di soglia (V_{th}) del 15% ogni generazione [2, 3].

Si sono ormai però raggiunti livelli che contrastano con i valori delle tensioni intrinseche dei singoli transistor, come il *bandgap* del silicio e il potenziale intrinseco delle giunzioni, per cui un'ulteriore diminuzione risulta molto difficoltosa. In Figura 1.2 [4] si può notare l'andamento dello *scaling* della tensione d'alimentazione nel tempo: il fatto importante da evidenziare è la diversa inclinazione delle due rette tracciate, a sottolineare come stia diventando sempre più difficile ridurre ulteriormente la tesione di alimentazione con il progredire della tecnologia.

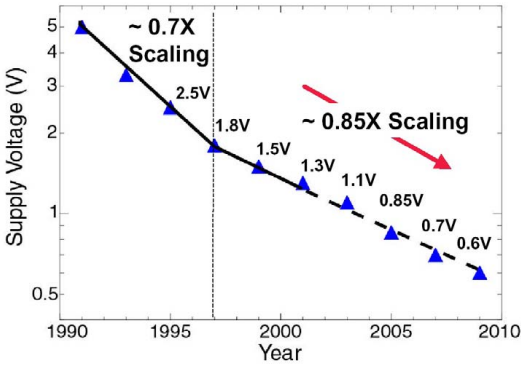


Figura 1.2: Scaling della Tensione di Alimentazione

Se da un lato però la riduzione di V_{DD} e V_{th} ha permesso di ottenere alte prestazioni e mantenere sotto controllo il consumo di *potenza dinamica*,

cioè il consumo dei transistor dovuto al cambio di stato (*switching activity*) delle porte logiche, dall'altro ha aumentato le correnti di perdita (*Leakage Currents*), prima fra tutte la *corrente di sottosoglia* (*Subthreshold Leakage Current*).

Con la progressiva riduzione della grandezza dei transistor si sono accentuati inoltre i cosiddetti *effetti di canale corto* (*Short Channel Effects SCEs*), che vanno ad influire negativamente sulla tensione di soglia e quindi sulla corrente di sottosoglia. Per evitare che questi effetti abbiano un'influenza eccessiva, si tende a ridurre sempre più lo spessore dello strato di ossido presente sopra al *gate* (*ossido di gate*): ciò però comporta un aumento della cosiddetta *corrente di perdita di gate* (*Gate Leakage Current*).

Dal grafico in Figura 1.3 [5] si può vedere come il contributo delle correnti di perdita al consumo del circuito sia adesso trascurabile e tenda a crescere sensibilmente, almeno in valore assoluto, se non in percentuale, rispetto alla potenza dinamica. Se andiamo ad analizzare l'andamento delle voci riguardanti la *leakage power* previsto per *System on Chip*, notiamo che per il primo periodo (2009-2015) sarà previsto un grande aumento nel consumo complessivo, che però si stabilizzerà nel secondo periodo (2016-2020).

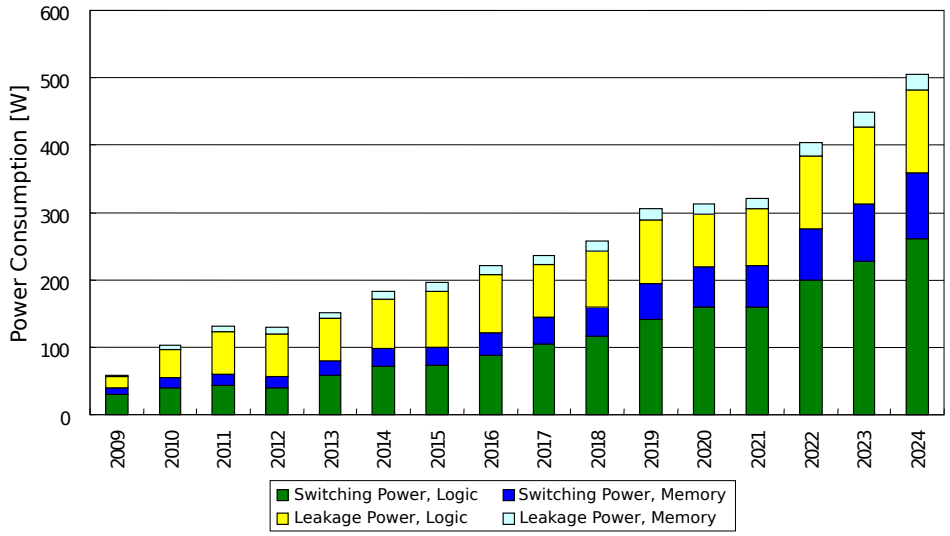


Figura 1.3: Trend del Consumo di Potenza Previsto per System on Chip

Il presente lavoro di tesi è suddiviso in 3 parti. Nel primo capitolo verranno analizzate le principali correnti di perdita e i fattori che ne influenzano l'intensità. Successivamente verranno illustrate le principali tecniche utili a ridurre il consumo di potenza, enfatizzando soprattutto quelle riguardanti il *design* circuitale. Infine verranno tratte alcune conclusioni.

2. Correnti di Perdita dei Transistor

In questa sezione verranno esaminate le principali correnti di perdita ed i fattori che maggiormente le influenzano. In Figura 2.1 [6] si possono notare le principali correnti di perdita:

- Corrente di sottosoglia (*Subthreshold Leakage*);
- Correnti di *gate* (*Gate Leakage*);
- Corrente della giunzione *drain/source-bulk* polarizzata in inversa (*Reverse Bias Source/Drain Junctions' Leakage*);
- Corrente di *punchthrough*;
- *Gate-Induced Drain Leakage*.

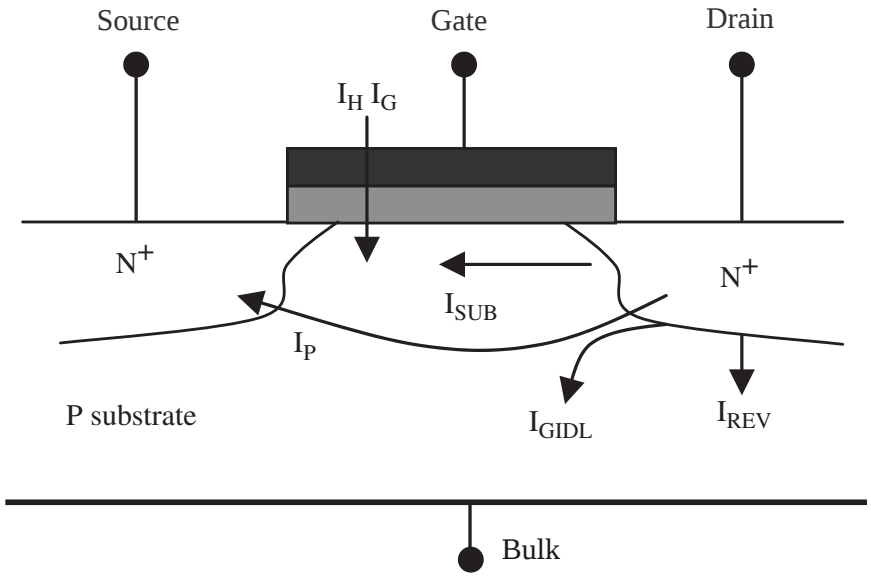
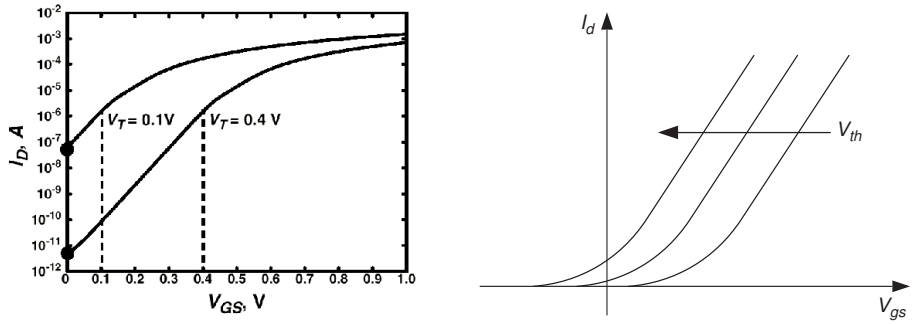


Figura 2.1: Correnti di Perdita in un Transistor NMOS

2.1 Corrente di Sottosoglia

La *corrente di sottosoglia* (I_{SUB}) è quella corrente che scorre fra il *drain* ed il *source* del transistor quando la tensione di gate V_{GS} è minore della tensione di soglia V_{th}, cioè quando il transistor è (idealmente) spento. In tali condizioni, le *correnti di diffusione* sono quelle che maggiormente contribuiscono alla corrente di sottosoglia.

Gli elementi che influenzano l'intensità di I_{SUB} sono molteplici, ma prima di tutto è importante sottolineare il legame che esiste fra la corrente di sottosoglia e la tensione di soglia: molti dei fattori che andremo a considerare, infatti, non influenzano direttamente la corrente di sottosoglia, bensì agiscono sulla tensione di soglia, comportando una variazione di I_{SUB}. Questo legame può essere meglio compreso analizzando i grafici di Figura 2.2a [7] e di Figura 2.2b [3].



(a) (b)

Figura 2.2: Corrente di Sottosoglia e Tensione di Soglia in un Transistor NMOS

Ciò che è rilevante è il fatto che abbassando la tensione di soglia, la corrente di sottosoglia (a $V_{GS} = 0$) aumenta esponenzialmente: come è stato detto nell'introduzione, ad ogni nuova generazione tecnologica CMOS (*nodo tecnologico*) per poter mantenere alte le prestazioni, a fronte di un abbassamento della tensione di alimentazione, si ha un abbassamento anche della tensione di soglia. Analizziamo quindi i principali elementi che influenzano la corrente di sottosoglia.

2.1.1 Temperatura

Molti circuiti integrati si trovano ad operare a temperature molto elevate, a causa soprattutto della dissipazione di potenza del circuito stesso: è molto importante quindi conoscere la relazione fra temperatura e corrente di sottosoglia.

In Figura 2.3[1] [2] è rappresentata la variazione della corrente di perdita totale del circuito (I_{OFF}): essendo una tecnologia CMOS a 0.35μm, la componente dominante di I_{OFF} è proprio la corrente di sottosoglia. Dal grafico si può capire che l'aumento della temperatura influenza la corrente di sottosoglia in due modi:

1. Fa aumentare linearmente S_t e quindi il transistor risulta sempre "meno spento";

2. Fa diminuire la tensione di soglia.

La sensibilità alla temperatura di V_{th} è stata misurata in circa $0,8$mV/°C.

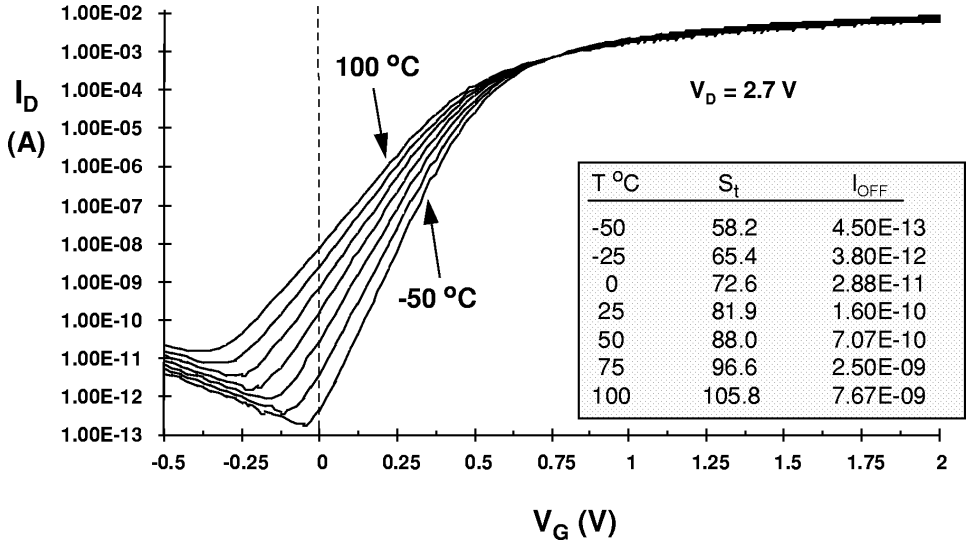


Figura 2.3: Effetto della Temperatura sulla Corrente di Sottosoglia

2.1.2 Lunghezza di Canale e V_{th} Rolloff

La tensione di soglia diminuisce con il diminuire della lunghezza di canale: tale fenomeno è noto come V_{th} *Rolloff* ed è ben descritto dal grafico in Figura 2.4a nella pagina successiva.

[1] Nel grafico viene usato il simbolo S_t (*subthreshold slope*) per indicare la pendenza in sottosoglia. Matematicamente è dato da:

$$S_t = \left(\frac{d(\log_{10} I_{ds})}{dV_{ds}} \right)^{-1} = 2,3 \frac{mkT}{q} = 2,3 \frac{kT}{q} \left(1 + \frac{C_{dm}}{C_{ox}} \right)$$

Fisicamente rappresenta l'efficacia con cui si riesce a spegnere un transistor quando V_{gs} è minore di V_{th}. Più piccolo è il suo valore, più agevole è lo spegnimento del transistor, cioè I_{ds} è minore.

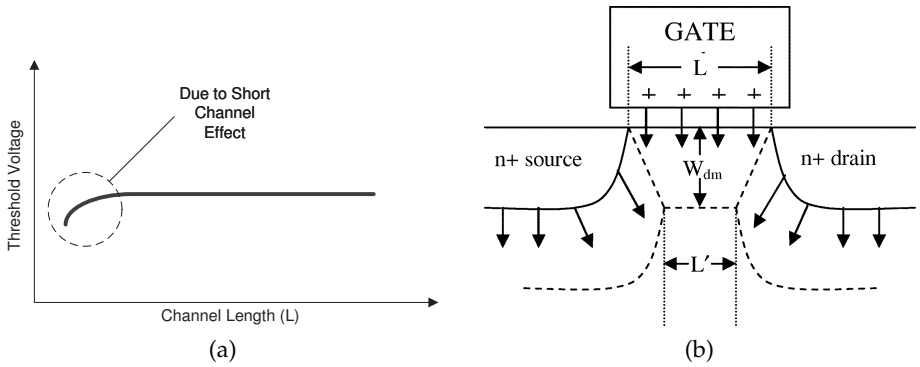


Figura 2.4: Effetto della Variazione della Lunghezza di Canale e V_{th} Rolloff

Tale diminuzione è dovuta al fatto che le regioni di svuotamento delle giunzioni *drain/source-bulk* penetrano maggiormente sotto al *gate*, in maniera non più trascurabile, come poteva essere nei dispositivi a canale lungo (vedi Figura 2.4b [2]): la tensione che si deve applicare al *gate* per ottenere l'inversione del canale è quindi inferiore, perché quest'ultimo è già in parte svuotato di portatori maggioritari.

2.1.3 Drain-Induced Barrier Lowering (DIBL)

Abbiamo visto che nei dispositivi a canale corto la larghezza delle regioni di svuotamento delle giunzioni *drain/source-bulk* non è più trascurabile. Tali regioni, però, possono arrivare anche ad interagire fra di loro abbassando il potenziale della barriera vista dal lato del *source*: applicando una grande tensione al *drain*, infatti, si ha un abbassamento della barriera di potenziale, con conseguente diminuzione della tensione di soglia e aumento della corrente di perdita fra *source* e *bulk* (corrente di *diffusione*).

2.1.4 Body Effect

In questo caso la tensione di soglia viene influenzata dal fatto che la polarizzazione inversa della giunzione *well-source* fa aumentare lo "svuotamento del *bulk*", comportando un aumento della tensione di soglia. Con l'aumentare del grado di inversione della giunzione, applicando ad esempio una tensione positiva al source, si ha un aumento della tensione di soglia. Il legame fra *body effect* e tensione di soglia (V_T) è riassunto nella seguente formula [8]:

$$V_T = V_{T0} + \gamma \left(\sqrt{|(-2)\phi_F + V_{SB}|} - \sqrt{|2\phi_F|} \right)$$

dove V_{T0} è la tensione di soglia del transistor in assenza di *body effect*, γ è chiamato *coefficiente dell'effetto body*, che quantifica l'effetto delle variazioni di V_{SB}, e ϕ_F è il *potenziale di Fermi*.

2.2 Correnti di Gate

In questa sezione sono descritte tutte le correnti che sono influenzate dalla progressiva riduzione dello strato di ossido di silicio sopra al *gate* (T_{ox}): queste correnti fino a pochi anni fa erano trascurabili rispetto alla corrente di sottosoglia, poiché lo spessore del suddetto strato era abbastanza spesso da limitarne l'intensità ($\geq 2\,\text{nm}$). Con la diminuzione della lunghezza di canale dei transistor, però, si è reso necessario diminuire ($\approx 30\%$ per generazione) anche T_{ox}, per poter limitare gli effetti di canale corto (*SCEs*): nelle tecnologie con T_{ox} inferiore ai 2 nm si è assistito ad un grande aumento delle correnti di *gate*. Ad esempio, per tecnologie con T_{ox} minore a $1,4\,\text{nm}$ le correnti di *gate* aumentano di 1000 volte alla successiva generazione, mentre la corrente di sottosoglia aumenta di 5 volte [3].

Oltre all'ulteriore consumo di potenza introdotto da queste correnti, un'altra conseguenza negativa è la perdita della classica assunzione riguardo l'impedenza d'ingresso infinita del transistor MOS, con ricadute sulle performance del circuito, dato che compare come resistenza in parallelo alla C_G (capacità di *gate*).

2.2.1 High-κ Gate Materials

La riduzione di T_{ox} serve innanzitutto a mantenere alta la corrente di saturazione: infatti, poiché tale corrente dipende linearmente da C_{ox} (capacità dello strato di ossido) [8]

$$I_{DSAT} = \mu_n C_{ox} \frac{W}{L} \left((V_{GS} - V_T)\, V_{DSAT} - \frac{V_{DSAT}^2}{2} \right)$$

si capisce come sia importante avere un'alto valore di C_{ox} per avere un'alta I_{DSAT}. La capacità dello strato di ossido è data, come tutte le capacità, da

$$C_{ox} = \frac{\kappa \varepsilon_0 A}{T_{ox}}$$

dove A è l'area dello strato di ossido e κ la costante dielettrica relativa dell'ossido. Si può agire quindi in tre modi per accrescere il valore di C_{ox}:

1. Ridurre T_{ox}, ma ciò comporta un aumento delle correnti di *gate*;
2. Aumentare l'area A, ma ciò è inconciliabile con il progresso della tecnologia;

3. Aumentare κ, cambiando l'ossido con un altro materiale.

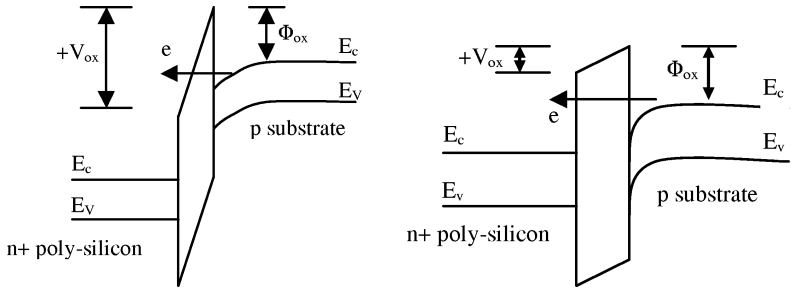
L'industria dei semiconduttori [9] negli ultimi anni ha iniziato a studiare nuovi materiali, basati soprattutto su *hafnio*, caratterizzati da una costante dielettrica maggiore di quella dell'ossido di silicio ($\kappa_{SiO_2} = 3,9$): tali materiali, chiamati *High-κ materials*, permettono quindi di avere alti valori di C_{ox} con uno spessore T_{ox} tale da non permettere il passaggio di alte correnti di *gate*.

2.2.2 Tunneling Dentro ed Attraverso l'Ossido di Gate

A causa dei sempre più piccoli valori di T_{ox}, il campo elettrico fra il *gate* ed il substrato risulta molto grande: ciò contribuisce a provocare il *tunneling* dei portatori (elettroni o lacune) attraverso lo strato di ossido (I_G). A questo fenomeno contribuiscono fondamentalmente due meccanismi:

1. *Fowler-Nordheim (FN) Tunneling*: in questo caso gli elettroni passano (*tunnel*) nella banda di conduzione dello strato di ossido di silicio (Figura 2.5a [2]);

2. *Direct Tunneling*: questo caso è presente quando si hanno strati molto sottili di ossido. Gli elettroni riescono quindi a passare (*tunnel*) direttamente al *gate* attraverso l'intervallo di energia proibito dell'ossido, senza entrare nella banda di conduzione dell'isolante (Figura 2.5b [2]).



(a) Fowler-Nordheim Tunneling (b) Direct Tunneling

Figura 2.5

2.2.3 Iniezione di Portatori Caldi dal Substrato all'Ossido di Gate

Come è stato detto sopra, la riduzione dello spessore di ossido ha comportato un innalzamento dell'intensità del campo elettrico attraverso lo stesso ossido: oltre a contribuire all'effetto *tunnel*, questo aumento

si accompagna con una crescita dei campi nel MOSFET. Gli alti campi elettrici nel transistor fanno sì che gli elettroni e le lacune riescano a guadagnare sufficiente energia per superare la barriera di potenziale ed entrare nell'ossido. Questo fenomeno (I_H) si chiama *iniezione di portatori caldi* (*injection of hot carriers*) ed è più probabile per gli elettroni che per le lacune, che hanno una *massa efficace* superiore. Alcuni dei portatori possono venire "intrappolati" nell'ossido alterando la tensione di soglia e quindi la corrente di sottosoglia.

2.3 Altre Sorgenti di Perdita

2.3.1 Corrente della Giunzione pn Polarizzata in Inversa

Solitamente le giunzioni *drain/source-well* sono polarizzate inversamente, quindi la corrente che scorre attraverso tali giunzioni (I_{REV}) ha principalmente due contributi:

- *Deriva* dei portatori minoritari;
- *Generazione* di coppie elettrone-lacuna nella regione di svuotamento.

Tale corrente è funzione dell'area della giunzione, del livello di drogaggio delle varie zone e del tasso di generazione legato alla difettosità del silicio.

2.3.2 Punchthrough

Come è stato precedentemente osservato, nei dispositivi a canale corto le grandezze delle regioni di svuotamento del *drain* e del *source* non sono più trascurabili rispetto alla lunghezza del canale. Se, in condizioni di transistor spento, aumentiamo di molto la tensione fra il *drain* ed il *source* (V_{ds}), otteniamo che le due regioni di svuotamento possono anche arrivare a fondersi nel *bulk* di silicio, ben distante dall'interfaccia silicio/ossido, se il nostro transistor è abbastanza piccolo: questo fenomeno è detto *punchthrough* (I_P). In questa condizione la barriera di potenziale dei portatori maggioritari del *source* è abbassata, in modo analogo al *GIDL* che vedremo di seguito, e alcuni di essi entrano nel substrato e vengono raccolti dal *drain*: ciò non fa altro che aumentare la corrente di sottosoglia.

2.3.3 Gate-Induced Drain Leakage (GIDL)

Questa corrente di perdita (I_{GIDL}) si verifica nei dispositivi a canale corto: per capirne la causa facciamo un parallelo idraulico. Immaginiamo che il *drain* ed il *source* siano due bacini idrici e che siano separati da una chiusa, la cui altezza dipende dal valore di V_{gs}.

Prendiamo in esame il caso di un transistor a canale lungo: la chiusa allora è una barriera alta e ben definita, che difficilmente si riesce ad oltrepassare. Quando viene applicata una tensione $V_{gs} \geq V_{th}$ l'altezza della chiusa si abbassa e raggiunge un valore tale per cui può iniziare a scorrere dell'acqua. Ovviamente bisogna applicare una tensione $V_{ds} \geq 0$ perché inizi a scorrere dell'acqua: in tal caso possiamo immaginare che il bacino del *drain* sia più in basso rispetto a quello del *source* e che si formi, quindi, una discesa che collega i due bacini per permettere all'acqua di scendere. Se invece fossimo nelle seguenti condizioni $V_{gs} \leq V_{th}$ e $V_{ds} \geq 0$ assisteremmo solamente ad un lieve abbassamento della chiusa, dovuto allo "sprofondare" del bacino del *drain*, ma non avremmo comunque un flusso di elettroni.

Nel caso di un transistor a canale corto invece, la chiusa che divide i due bacini non è molto spessa e in determinate condizioni può esserci passaggio di elettroni, senza che tuttavia il transistor sia acceso. Se infatti consideriamo $V_{ds} \geq 0$ e $V_{gs} \leq V_{th}$ abbiamo che il bacino del *drain* si abbassa rispetto a quello del *source* comportando un leggero abbassamento della chiusa: tuttavia la barriera è molto più sottile rispetto a prima e quindi si può assistere ad una passaggio di elettroni. Ciò è dovuto al fatto che non tutti gli elettroni hanno lo stesso livello di energia; ce ne sono alcuni che presentano un'energia maggiore degli altri, sufficiente ad oltrepassare la chiusa, secondo la distribuzione statistica propria delle energie degli elettroni. Classicamente, tale distribuzione è quella di *Maxwell-Boltzmann*[2].

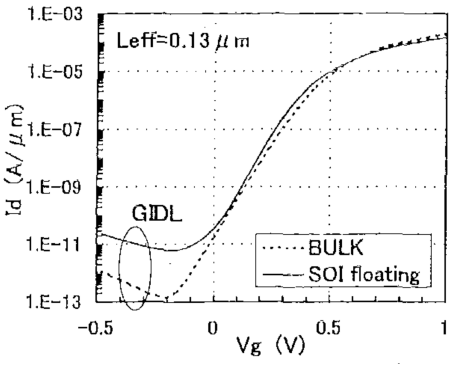


Figura 2.6: Caratteristica I_d-V_g

L'andamento dell'effetto del GIDL è riportato in Figura 2.6 [10]. Si può notare come la corrente di *drain* diminuisca in maniera molto accentuata a partire da 0,6-0,7 V (tensione di soglia) fino a 0 V nella tensione di sottosoglia; a tensioni di *gate* ancora minori invece ricomincia a crescere a causa del GIDL.

[2]La distribuzione di Maxwell-Boltzmann è una funzione di distribuzione riguardante le particelle in un sitema che obbedisce alle leggi della fisica classica. Tale distribuzione descrive la probabilità che una particella abbia un'energia (o una velocità) compresa fra $E + dE$ (o $v + dv$). Ci sono, tuttavia, delle ipotesi che il sistema deve soddisfare affinché sia possibile applicare tale ragionamento: le particelle devono essere *distinguibili*, il sistema *lineare*, *isotropo* ed i processi statistici alla base dello stato del sistema devono essere *processi markoviani*. Se ad esempio andiamo a considerare un sistema in *meccanica quantistica*, allora la prima ipotesi decade e si manifestano delle diverse distribuzioni, ovvero quella di *Fermi-Dirac* e quella di *Bose-Einstein*.

3. Tecniche per la Riduzione del Consumo

In questa sezione saranno esposte le principali tecniche di riduzione delle correnti di perdita. Verranno esposte essenzialmente le tecniche legate al *design* del circuito, mentre saranno tralasciate quelle più tecnologiche, inerenti alla costruzione ottimale del singolo componente.

3.1 Transistor Stacks

Questa tecnica si basa su una semplice considerazione: in uno *stack*[1] di MOSFET si ottiene una riduzione maggiore della corrente di sottosoglia quando più di un componente è spento, rispetto al caso in cui ve ne sia solo uno spento. Per analizzare quanto appena detto, prendiamo in esame una porta NAND CMOS a due ingressi (Figura 3.1).

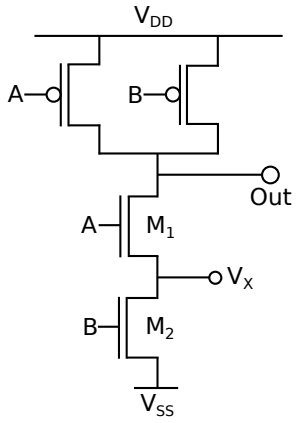


Figura 3.1: Nand CMOS

Poniamo che M_1 e M_2 siano entrambi spenti ($V_g = 0$): la tensione al nodo interno della PDN (*Pull Down Network*), V_X, è positiva a causa della debole corrente di *drain* che circola attraverso M_2. Ciò comporta varie conseguenze:

- La tensione *gate-source* V_{gs} di M_1 diventa negativa, riducendo la corrente di sottosoglia (vedi il grafico in Figura 2.2b a pagina 6);
- In riferimento all'effetto body (esposto in 2.1.4 a pagina 8) abbiamo visto che una tensione positiva al source (di M_1) comporta un aumento della tensione di soglia, con conseguente diminuzione della corrente di sottosoglia;
- La tensione drain-source V_{ds} di M_1 diminuisce, comportando una diminuzione del DIBL, analizzato in 2.1.3 a pagina 8.

[1] Per *stack* si intende una struttura circuitale serie fra alimentazione (V_{DD}) e massa (V_{SS}).

Molto importante risulta quindi la configurazione degli ingressi: l'obiettivo, infatti, è quello di trovare il *minimum input vector*, cioè quell'insieme di valori degli input che permettono di massimizzare il numero di dispositivi spenti quando il circuito è messo in *standby*.

Per trovare tale "vettore", l'approccio più intuitivo risulta quello di analizzare (attraverso simulazioni) tutte le possibili combinazioni degli ingressi. Un tale algoritmo, che si basa su uno schema di tipo *brute force*, è caratterizzato però da una complessità esponenziale: dato infatti un circuito formato da n input, si ottengono 2^n casi possibili da analizzare. Se ne deduce che si può agire in questo modo solo se si sta trattando con circuiti con pochi input. In caso contrario si deve utilizzare un metodo probabilistico, come esposto in [11]. In [12] invece viene illustrata una particolare euristica per poter individuare il *minimum input vector*.

3.1.1 Stacking Transistor Insertion

Sempre in [12] viene esposto un ulteriore processo da eseguire dopo aver trovato il *minimum input vector*: si tratta di inserire un transistor (*leakage control transistor*) nelle porte in cui non si è riusciti ad operare con l'*input vector*, cosicché anche queste porte possano avere una bassa corrente di perdita. Un esempio dell'inserimento di tali transistor in una porta NAND è illustrato in Figura 3.2.

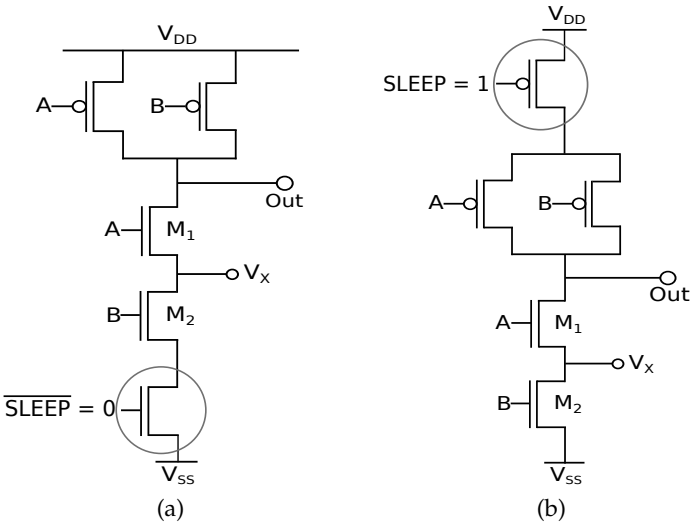


Figura 3.2: NAND con Due Diverse Inserzioni del Leakage Control Transistor

L'inserimento viene eseguito in serie fra i transistor della porta ed il *ground* (Figura 3.2a) o l'alimentazione (Figura 3.2b). Un segnale di *sleep*

ne controlla l'accensione e lo spegnimento. L'aggiunta di tale transistor tuttavia va a provocare un aumento del consumo e del ritardo della porta.

Implementazione Circuitale per la Memorizzazione dell'*Input Vector*

In Figura 3.3 [13] è mostrata la realizzazione circuitale di due *latch* utili alla memorizzazione dei valori dell'*input vector*.

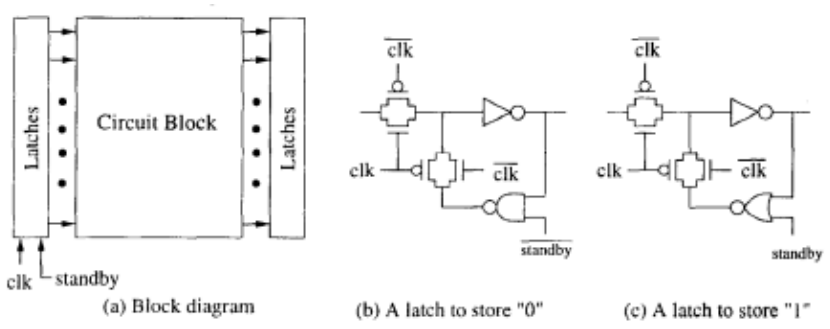


Figura 3.3: Memorizzazione del Minimum Input Vector

In entrambi i casi il valore memorizzato è quello presente all'uscita dell'*inverter*. Il funzionamento globale prevede che il valore che può essere memorizzato dal latch, venga tenuto in memoria e trasmesso ai vari transistor che devono essere comandati, una volta che il segnale di *standby* commuta ad "1".

Prendiamo in esame il primo *latch*, quello per lo *store* dello "0" e poniamo che il valore memorizzato sia quello presente dopo l'*inverter* (ingresso della NAND). Il funzionamento è divisibile in due parti:

standby = 0 Il circuito è in modalità attiva e siamo in memorizzazione quando *clk* = 0, poiché il *T-gate* in alto è interdetto, mentre quello che chiude l'anello di retroazione è attivo; viceversa, quando *clk* = 1, il *T-gate* in basso è spento, con conseguente interruzione della retroazione e possibiltà di scrittura tramite il *T-gate* in alto che è attivo. Poiché *standby* è "0", la NAND in basso a destra nega a sua volta il segnale che passa attraverso l'*inverter*: se ad esempio viene memorizzato "0", all'input della NAND arrivano lo "0" in questione e l'"1" dello *standby* negato, determinando un'uscita pari ad "1".

standby = 1 Il circuito è in *standby* e l'unico valore che può essere tenuto in memoria è uno "0". Supponiamo infatti che ci sia un "1" all'uscita dell'*inverter*: all'ingresso della NAND si presenteranno questo "1" e lo "0" dello *standby* negato, comportando un "1" in uscita e quindi uno "0" dopo l'*inverter*, che dunque si mantiene come unico stato possibile.

Il funzionamento del *latch* che memorizza "1" è speculare e l'utilizzo della NOR al posto della NAND serve a far sì che, quando *standby* è "1", all'uscita del circuito ci sia sicuramente "1": questo è dato dal fatto che l'uscita di una NOR è sempre 0 se all'ingresso è presente almeno un "1".

Risultati

La riduzione della corrente di perdita ottenibile varia in base al circuito a cui viene applicata tale tecnica. In [13], ad esempio, viene analizzato un circuito (un particolare *adder* CMOS realizzato in tecnologia 0.1 μm) per il quale la riduzione è pari al doppio di quella che si ottiene spegnendo meno transistor. In Figura 3.4 [14] vengono riportati ulteriori risultati riguardanti altri circuiti.

Questa tecnica prevede infine minime maggiorazioni (*overheads*) per quanto riguarda area e consumo di potenza.

Unit	Leakage Reduction (%)	Area Increase (%)	Min. idle time (us)
32-bit CLA	64.84	0.92	25.16
8x8 Multiplier	21.00	0.13	112.36
8-bit Static Adder	95.22	2.05	10.52
32-bit Shifter	79.13	0.27	22.40
15-bit CLA	66.02	0.96	46.66
3-to-1 Mux	95.19	1.65	27.77
32 2-input XOR	24.11	6.37	20.01
32 2-input NAND	93.81	9.37	0.70
32 2-input AND	98.53	8.22	0.39
32 2-input NOR	97.71	6.87	0.11
32 2-input OR	98.82	5.46	0.52

Figura 3.4

3.2 Multiple V_{th} Design

Come è stato detto nell'introduzione, con l'avanzare della tecnologia si è sempre teso a diminuire la tensione di alimentazione al fine di tenere sotto controllo il consumo di potenza dinamica: se guardiamo infatti l'equazione che la descrive [8]

$$P = C_L V_{dd}^2 f$$

in cui C_L rappresenta la capacità di carico, V_{dd} la tensione di alimentazione ed f la frequenza a cui opera il circuito, notiamo una dipendenza quadratica della potenza da V_{dd}. Ciò spiega perché sia assai efficacie abbassare la tensione di alimentazione per diminuire il consumo di potenza.

Il lato negativo di questo *scaling* è l'aumento del ritardo (t_p) della porta: si può vedere ciò nella seguente relazione [15]

$$t_p \propto \frac{C_L V_{dd}}{I_{DS}} \simeq \frac{C_L V_{dd}}{A(V_{dd} - V_{th})^2}\ , \qquad A = \text{constant}$$

in cui una riduzione di V_{dd} fa diminuire il denominatore molto più velocemente del numeratore, causando un aumento del valore di t_p.

Un modo per evitare di incrementare il ritardo è quello di diminuire la tensione di soglia V_{th} dei dispositivi: tuttavia ciò provoca una crescita della corrente di sottosoglia e del GIDL, come spiegato in 2.1 a pagina 6.

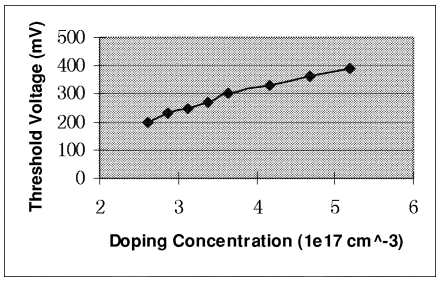
In questa sezione analizzeremo varie tecniche di soppressione delle correnti di perdita accomunate tutte dall'utilizzo di transistor sia a bassa che ad alta tensione di soglia (*low-threshold/high-threshold transistors*) all'interno dello stesso chip, da cui il nome *Multiple* V_{th}. L'utilità di questo doppio V_{th} risiede nel fatto che gli *high-threshold transistor* sono caratterizzati da una bassa corrente di sottosoglia, mentre i *low-threshold transistor* permettono di mantenere alte le prestazioni in termini di ritardo.

Per costruire tali transistor esistono vari tecniche:

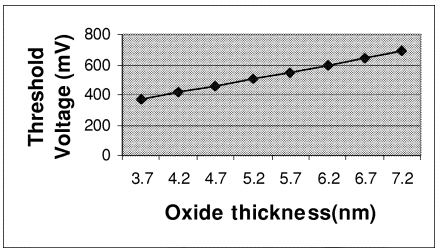
1. *Multiple Channel Doping*. Per ottenere diverse tensioni di soglia vengono drogati in maniera diversa i canali dei transistor. La dipendenza della tensione di soglia dalla concentrazione del doping è visibile in Figura 3.5a nella pagina successiva.

2. *Multiple Oxide CMOS* (M_{ox}CMOS). In tale tecnica, crescendo strati di ossido di spessore diverso sotto al gate, si ottengono diverse tensioni di soglia. In Figura 3.5b nella pagina seguente si può notare l'andamento della tensione di soglia rispetto allo spessore dell'ossido di gate. Negli *high-threshold transistor* viene quindi applicato uno strato maggiore di ossido: ciò, oltre a portare una diminuzione della corrente di sottosoglia, produce il vantaggio di diminuire le correnti di

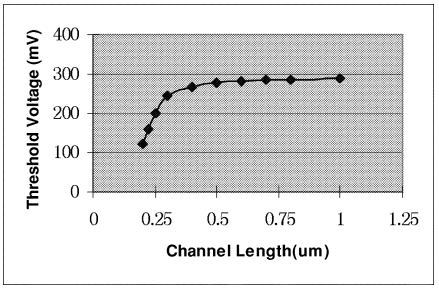
perdita di gate (vedi 2.2) ed il consumo dinamico di potenza, poiché le capacità di gate si riducono. Tuttavia in dispositivi a dimensioni molto ridotte, l'aumento dello strato di ossido comporta un aumento degli effetti di canale corto (*SCEs*): per mantenere sotto controllo questi effetti si tende, quindi, ad aumentare anche la lunghezza di canale.



(a) V_{th} e Densità del Doping del Canale (Tecnologia 0.25μm)



(b) V_{th} e Spessore dell'Ossido di Gate (Tecnologia 0.25μm)



(c) V_{th} Rolloff (NMOS)

Figura 3.5

3. *Multiple Channel Length*. Nei transistor a canale corto la tensione di soglia diminuisce al diminuire della lunghezza di canale (V_{th} *Rolloff*), come precedentemente spiegato in 2.1.2 a pagina 7. L'andamento della tensione di soglia rispetto alla lunghezza di canale può essere visto in Figura 3.5c. Poiché però per avere *high-threshold transistor* bisogna aumentare la lunghezza di canale, si hanno effetti negativi sul ritardo ed il consumo di potenza, a causa dell'aumento delle capacità di gate.

4. *Multiple Body Bias*. Esistono delle tecniche che invece di utilizzare direttamente *high* o *low threshold transistor*, polarizzano, tramite l'applicazione di determinate tensioni, il *body* del transistor, in modo da alterare la tensione di soglia del singolo componente (vedi effetto body 2.1.4 a pagina 8). Il problema è che, se su una stessa *well* sono

impiantati più transistor, non si può applicare una diversa polarizzazione ai singoli transistor. Per un'applicazione estensiva di tale tecnica, si rivela fondamentale l'utilizzo di transistor di tipo SOI, *Silicon-On-Insulator*. La struttura del *bulk* è composta da un triplo strato, in ordine silicio-isolante-silicio. Come si può vedere in Figura 3.6 [16], il *bulk* del transistor, che si limita quasi al solo canale, è isolato da quello degli altri. I vantaggi che questa tecnologia apporta sono molteplici: ad esempio le capacità parassite al *drain* e al *source* si riducono poiché, se consideriamo un NMOS, l'area di contatto fra le zone *n* (*drain* e *source*) e quelle *p* diminuisce a causa della presenza dell'isolante.

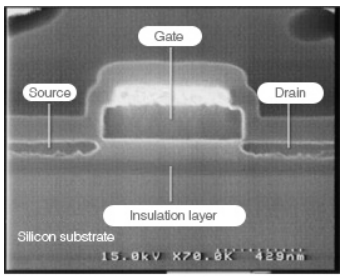


Figura 3.6: MOSFET SOI

Analizziamo ora diverse tecniche di riduzione delle correnti di perdita che utilizzano *multi-threshold transistor*.

3.2.1 Multithreshold-Voltage CMOS

Come detto in precedenza, questa è una delle tecniche, chiamata *Multithreshold-Voltage CMOS* (MTCMOS), che impiega transistor con due differenti tensioni di soglia: nello specifico gli *high-threshold transistor* collegano all'alimentazione e al *ground* i *low-threshold transistor* che compongono la porta.

In Figura 3.7 sono illustrate alcune implementazioni di una porta NAND con la tecnica MTCMOS [15]. I terminali della porta non sono collegati direttamente a V_{dd} ed a GND, bensì a due linee "virtuali" $V_{dd}V$ e $GNDV$, poiché un PMOS ed un NMOS (entrambi ad alta tensione di soglia) sono frapposti rispettivamente fra V_{dd} e il PUN e/o fra GND ed il PDN. Un segnale di *sleep* (SL) controlla l'accensione e lo spegnimento degli *high-threshold transistor* nel caso in cui la porzione di circuito in cui si trova la porta sia, rispettivamente, in attività o in *standby*.

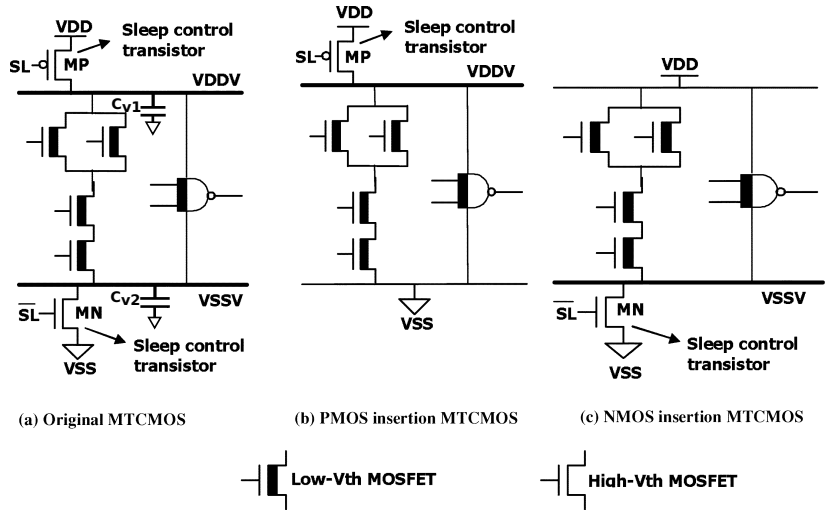


Figura 3.7: Alcuni Schemi di Porte in MTCMOS

Funzionamento

In riferimento alla Figura 3.7a, in modalità attiva il segnale di *sleep* è fissato a "0" ed i transistor MP e MN sono accesi: poiché le resistenze parassite sono basse, grazie ad un opportuno dimensionamento dei MOS, le linee virtuali $V_{dd}V$ e $GNDV$ sono prossime, come valore di tensione, alle linee reali V_{dd} ed GND. La porta NAND può quindi operare normalmente, avendo anche alte *performance* in termini di ritardo in quanto sono impiegati transistor a bassa tensione di soglia.

In modalità *sleep* invece, il segnale SL è alto, con MP e MN che sono spenti. Le linee virtuali sono sconnesse dall'alimentazione e dal *ground* e sono quindi flottanti. L'attuale porta NAND, a differenza della sua implementazione in CMOS statica, non è percorsa da un'alta corrente di sottosoglia, poiché MP e MN provvedono a sopprimerla: essendo infatti *high-threshold transistor*, essi sono percorsi da una corrente di sottosoglia molto minore rispetto a quella che scorrerebbe nei *low-threshold transistor* collegati direttamente alle linee reali. Il consumo di potenza in condizioni di standby viene quindi di molto ridotto.

Per tale tecnica non è richiesta una grande maggiorazione nell'utilizzo di area, poiché i transistor aggiunti per creare le linee virtuali possono essere condivisi fra più porte: non devono inoltre essere aggiunti moduli di controllo o di altra specie, a parte il generatore del segnale SL.

Analisi delle Prestazioni

Nell'analisi che ci prestiamo a fare andremo a considerare principalmente due fattori: la *grandezza* dei dispositivi MP e MN e le *capacità* C_{V1} e C_{V2}. Se infatti si aumenta il fattore di forma W/L dei due transistor si ottiene una *on-resistance* più piccola. Per quanto riguarda invece le capacità C_{V1} e C_{V2}, queste vanno a contribuire al ritardo totale della porta.

In Figura 3.8 [15] è mostrata la dipendenza esistente fra il ritardo della porta (*gate delay time*) e la tensione efficace V_{eff}[2] nei confronti della larghezza degli *high-threshold transistor* normalizzata rispetto a quella dei *low-threshold transistor* (W_H/W_L).

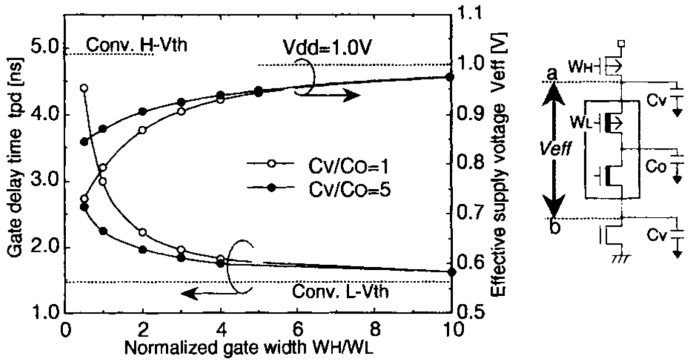


Figura 3.8: Ritardo della Porta e di V_{eff} vs Grandezza Normalizzata del Gate

Si può notare come alti valori di C_V e di W_H (ad esempio $C_V/C_O = 5$ e $W_H/W_L = 5$, dove C_O è la capacità al nodo d'uscita) comportino

[2] V_{eff} è definita come il valore minimo della differenza fra le tensioni $V_{dd}V$ e $GNDV$.

alte prestazioni (valori bassi di ritardo) e una tensione di alimentazione per le porte interne simile a quella che sarebbe data dalle linee reali ($V_{eff} \simeq V_{dd}$, $V_s \simeq 0$).

Dimensionamento nei Circuiti MTCMOS

Un adeguato dimensionamento dei transistor coinvolti in un circuito MTCMOS può portare ad ottenere prestazioni migliori senza eccedere nel consumo d'area. Da un lato, infatti, un impiego di transistor molto grandi comporta un consumo eccessivo d'area e maggiore energia per eseguire lo *switch* del dispositivo, mentre nel caso in cui vengano impiegati transistor piccoli, il circuito risulterà più lento a causa della *on-resistence* più alta.

Consideriamo un'implementazione MTCMOS come in Figura 3.7c e 3.9: quando SL è a 0, l'NMOS di controllo è acceso ed è quindi approssimabile con una resistenza lineare, detta *on-resistance* (vedi Figura 3.9 [17]). Quando l'uscita dell'*inverter* commuta da "1" a "0", le cariche della capacità d'uscita C_{load}, dopo aver percorso il transistor M_2, passano attraverso la *on-resistance* R: ciò induce una tensione al nodo X (V_X) e quindi un aumento della tensione di soglia di M_2 per *effetto body*.

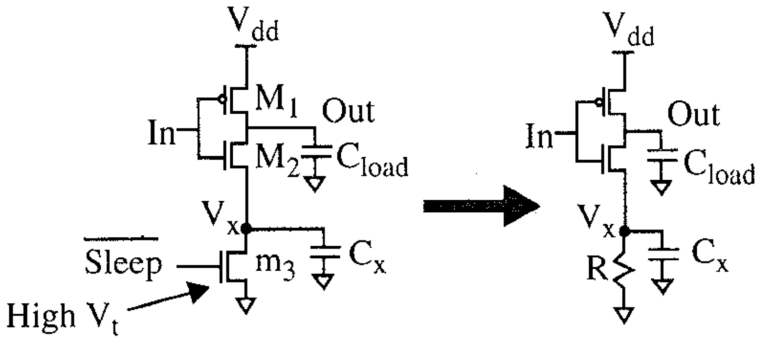


Figura 3.9: Modellizzazione del Transistor di Sleep come una Resistenza

Di conseguenza la corrente di scarica diminuisce in intensità (anche V_{ds} di M_2 si riduce) e ciò comporta un maggior tempo di scarica dell'uscita. Per ridurre V_X, e quindi limitare l'effetto negativo sul ritardo, si deve aumentare la grandezza W/L del transistor di *sleep*: tale aumento tuttavia è limitato dall'affacciarsi di altri problemi come l'accrescimento della corrente di sottosoglia, l'aumento di dissipazione di potenza di *switch* ed il maggior consumo d'area.

L'utilizzo della tecnica MTCMOS può comportare all'interno di un blocco logico delle situazioni di conduzione inversa, che si realizza quando il valore dell'output di una porta viene influenzato, non dal segnale di

input, ma da tensioni presenti in altri nodi del circuito. Questo fenomeno viene illustrato in Figura 3.10 [17]. All'inizio tutti gli *inverter*, tranne quello a sinistra, hanno il nodo d'uscita a "1" (l'input è "0"). Successivamente gli input dei due *inverter* più a destra commutano, facendo scaricare i nodi d'uscita, in una sorta di "*charge sharing*" transitorio: a questo punto una parte della carica totale contenuta nelle due capacità del nodo d'uscita scorre direttamente attraverso l'NMOS di *sleep*, facendo alzare la tensione della linea virtuale ad un valore V_X.

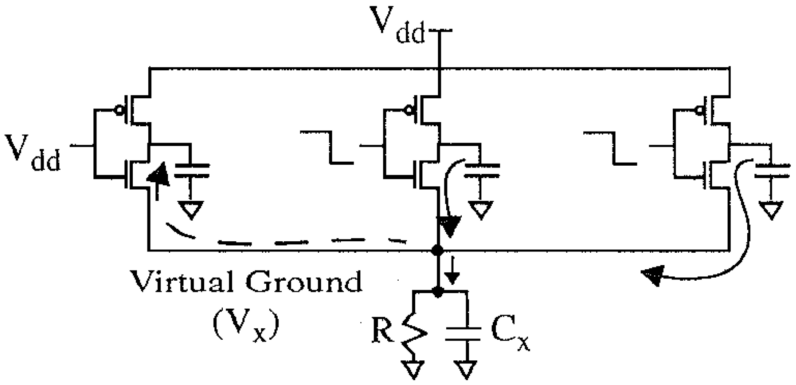


Figura 3.10: Conduzione Inversa nella MTCMOS

Essendo il nodo di output del primo *inverter* a zero, si viene a formare una differenza di tensione fra questo nodo e la *virtual line* e quindi una parte delle cariche andrà a finire nella capacità d'uscita del primo *inverter*. Ciò ha come effetto quello di rendere la porta più veloce, in quanto la tensione V_X è più bassa di quella che si avrebbe nel caso in cui tutta la carica scorresse nel'NMOS di *sleep*. Inoltre, se consideriamo la capacità di output dell'*inverter* più a sinistra, la commutazione basso-alto avverrà più velocemente in quanto il nodo d'uscita è già in parte carico, se si verificherà un'adeguata commutazione dell'ingresso. Ciò però comporta lo svantaggio di ridurre i margini di rumore della porta.

Implementazione di un Latch in MTCMOS

Un'attenzione particolare deve essere rivolta alla realizzazione di *latch* in modalità MTCMOS: come si sa, il *latch* è un'unità che serve alla memorizzazione di un valore ed è quindi importante riuscire a tenere memorizzata questa informazione, anche quando il circuito è in modalità *sleep*, con le linee flottanti.

In Figura 3.11a [15] è visibile un *latch* realizzato in MTCMOS: gli *inverter* G_2 e G_3 sono realizzati con *high-threshold transistor* e sono connessi

all'alimentazione ed al *ground* reali. L'*inverter* G_1 invece è realizzato in MTCMOS seguendo lo schema di Figura 3.7a.

La differenza con un *latch* costruito in CMOS statica risiede in due parti:

- I *T-gate* e l'*inverter* G_1 sono costruiti con *low-threshold transistor*;
- È presente un *inverter* aggiuntivo (G_3) che serve per tenere memorizzato il dato anche quando si è in modalità *sleep*.

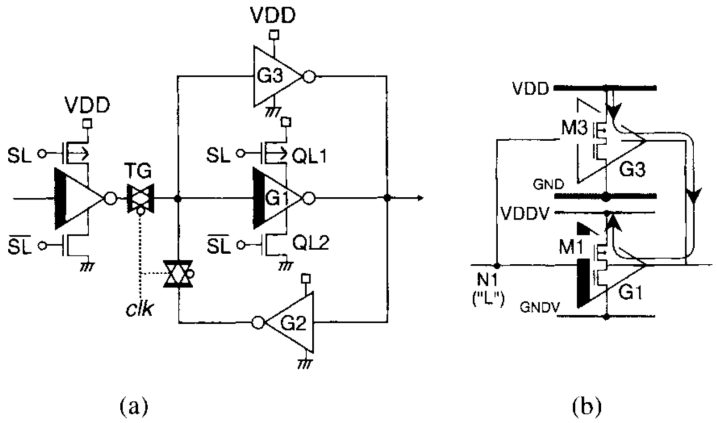


Figura 3.11: Latch MTCMOS

La maggior velocità è quindi data dalla combinazione fra i *T-gate* e l'*inverter* G_1 costruiti con transistor a bassa tensione di soglia. Il funzionamento può essere così diviso in due parti:

SL=0 Modalità Attiva: il circuito si comporta come un classico *latch*, che qui risulta essere formato dai due *T-gate* e gli *inverter* G_1 e G_2, quest'ultimo usato in retroazione. Al variare del *clock* quindi ci troviamo o in fase di memorizzazione (CLK=1) o di scrittura (CLK=0); come è stato detto, la memorizzazione risulta più veloce grazie all'impiego di *low-threshold transistor* nel *T-gate* e nell'*inverter* G_1;

SL=1 Modalità Sleep: in tale modalità l'unico problema risulta essere quello di tenere memorizzato il dato. Se infatti il nostro circuito si limitasse agli *inverter* G_1 e G_2 questa operazione risulterebbe problematica a causa del fatto che G_1 risulta flottante. L'aggiunta di G_3 risolve questo problema: in modalità *sleep* il *latch* risulta quindi formato dagli invertitori G_3 e G_2; quest'ultimo fa sempre parte dell'anello di retroazione.

Per costruire il *latch* abbiamo utilizzato il *design* della MTCMOS della Figura 3.7a: ci si può chiedere quindi perché non impiegare le altre configurazioni illustrate in Figura 3.7, dato che comportano un minore utilizzo di transistor. Ciò è illustrato dalla Figura 3.11b: se ad esempio collegassimo G_1 direttamente all'alimentazione virtuale senza frapporre un transistor, in modalità *sleep* protrebbe crearsi un cammino conduttivo fra l'alimentazione di G_3 e l'alimentazione virtuale di G_1, provocando un eccessivo consumo di potenza in regime di *standby*, dato che esiste una differenza di potenziale fra le due linee. L'utilizzo quindi di entrambi i transistor di *sleep* risulta fondamentale per evitare questo effetto.

3.2.2 Super Cut-off CMOS

Nella tecnica appena analizzata (MTCMOS) abbiamo visto che vengono utilizzati transistor sia a bassa tensione di soglia (0,2-0,3 V) che ad alta tensione di soglia (0,6-0,7 V). Di conseguenza il valore della tensione di alimentazione V_{dd} non può scendere molto al di sotto di 1 V, poiché ciò non permetterebbe la commutazione dei transistor ad alta tensione di soglia. La tecnica che stiamo per andare ad analizzare, denominata *super cut-off CMOS* (SCCMOS), impiega esclusivamente *low-threshold transistor*: essa permette di utilizzare una tensione di alimentazione inferiore a 0,5 V e transistor con una tensione di soglia minore di 0,2 V, senza tuttavia andare ad intaccare le prestazioni e senza aumentare le correnti di perdita [18].

In Figura 3.12 [18] è illustrato il *design* fondamentale che caratterizza questa tecnica:

- M_1 è l'equivalente del transistor di *sleep* della MTCMOS (vedi 3.2.1 a pagina 20): infatti è inserito fra l'alimentazione e la porta per creare una linea virtuale (V_{ddV});

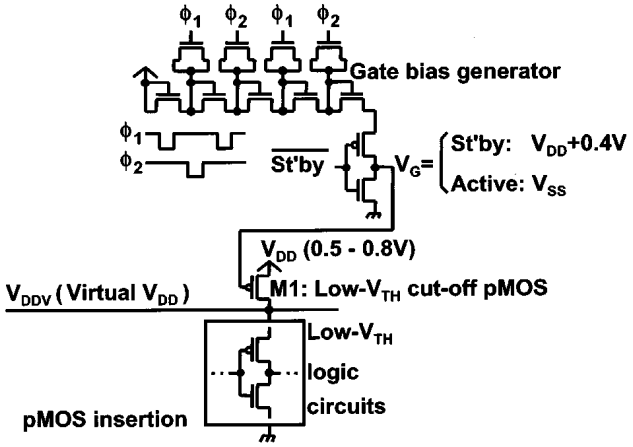


Figura 3.12: Super Cut-off CMOS

- Il segnale che regola l'accensione e lo spegnimento di M_1 è generato da un *inverter*, su comando di un segnale di *standby* (l'equivalente segnale di *sleep* della MTCMOS). L'output però, assume come valore "alto" $V_{dd} + 0,4V$: ciò è realizzato tramite un *gate bias generator* collegato al *source* del PMOS dell'*inverter*. L'utilità di avere una tensione maggiore di quella fornita dall'alimentazione risiede nel fatto che essa permette di spegnere in maniera più efficacie M_1, dato che è un *low-threshold transistor*, e quindi di limitare maggiormente la corrente di sottosoglia. In modalità attiva invece ad M_1 viene applicata V_{ss}, ovvero *ground*, sempre tramite l'*inverter*.

3.2.3 Dual Threshold CMOS

Un circuito digitale è assimilabile ad un grafo dove i nodi sono le varie porte logiche e gli archi le varie interconnessioni. Ogni porta logica è caratterizzata da un ritardo fra l'input e l'output: se prendiamo un percorso attraverso il grafo, questo sarà caratterizzato da una maggiore o minore velocità rispetto ad un altro, a seconda del ritardo delle porte che attraversa. Il percorso che determina la velocità massima del circuito è detto *percorso critico* (*critical path*). Ciò non significa tuttavia che tutti gli altri siano *non-critical path*, poiché può esistere più di un percorso critico.

Appare quindi logico pensare di realizzare le porte appartenenti ai percorsi critici con transistor a bassa tensione di soglia (per poter avere migliori risultati in termini di ritardo), mentre i transistor che riguardano percorsi non critici possono essere ad alta tensione di soglia (per poter limitare le correnti di perdita, anche a scapito delle prestazioni). La Figura 3.13 [2] spiega molto bene l'idea di fondo di questa tecnica.

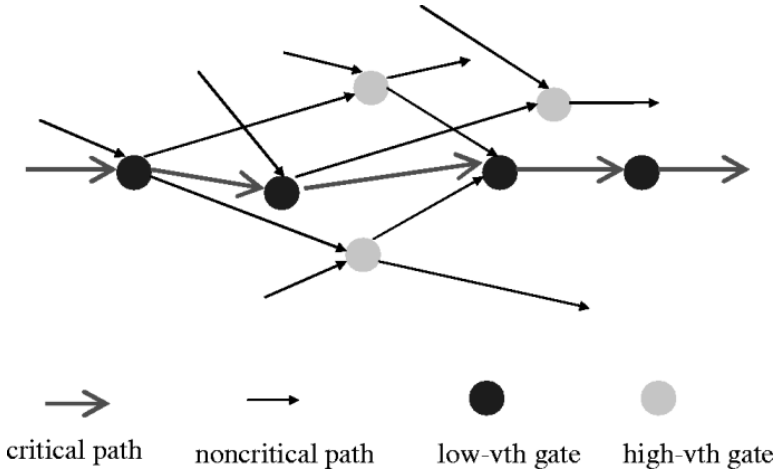


Figura 3.13: Dual Threshold CMOS

Tuttavia è bene ricordare che certi *non-critical path* possono diventare percorsi critici: di conseguenza in un percorso non critico non necessariamente tutti i transistor sono *high-threshold transistor*.

La parte più importante di questa tecnica riguarda il modo di trovare i possibili percorsi critici all'interno del circuito. Inoltre è fondamentale trovare i transistor a cui assegnare un'alta tensione di soglia: come è stato detto infatti, può accadere che alcune porte appartenenti ad un percorso non critico possano comunque ad un certo punto essere coinvolte in un *critical path*. Risulta infine importante anche trovare un'*adatta* alta tensione di soglia: a livello di circuito, infatti, vengono richieste specifiche a livello di ritardo che costringono ad utilizzare solo certi valori per la tensione

di soglia degli *high-threshold transistor*. Un algoritmo che svolge quanto spiegato finora lo si trova in [19].

A differenza delle tecniche MTCMOS e SCCMOS, la *Dual Threshold CMOS* presenta il vantaggio di poter essere utilizzata anche in condizioni di funzionamento attivo del circuito: infatti, le due tecniche precedentemente prese in considerazione sono impiegate solo in regime di *standby* per poter limitare la corrente di sottosoglia, mentre ciò che abbiamo appena analizzato consente di limitare le correnti di perdita anche quando il circuito è attivo senza andare ad intaccare le prestazioni.

Un altro vantaggio risiede nel fatto di non richiedere transistor aggiuntivi, a differenza della MTCMOS e della SCCMOS che prevedono l'utilizzo dei transistor di *sleep*. Di conseguenza non c'è *overhead* di area.

3.2.4 Variable Threshold CMOS

In questa tecnica, come nella successiva, non vengono impiegati transistor ad alta e a bassa tensione di soglia, bensì si va ad agire tramite determinati contatti sul *body* (polarizzazione), in modo da variarne la tensione e quindi influire sulla tensione di soglia. Come già spiegato alla fine di 3.2, i transistor utilizzati in questa operazione sono quelli basati sulla tecnologia *Silicon-On-Insulator*.

La *Variable Threshold CMOS* (VTCMOS) può essere analizzata come un parallelo della MTCMOS: anche in questa tecnica infatti si individuano due periodi di funzionamento (modalità attiva e *standby*), regolati da opportuni segnali. La differenza sta nel fatto che mentre nella MTCMOS i segnali di *sleep* sono gli input di transistor, qui i "segnali" in questione vanno ad agire sulla tensione del *bulk* dei transistor che compongono la porta.

Prendiamo quale esempio principe l'*inverter* di Figura 3.14, come riportato in [20]. In modalità attiva, viene applicata al *body* dei due transistor una tensione che serve solamente a tenere sotto controllo le fluttuazioni della tensione di soglia.

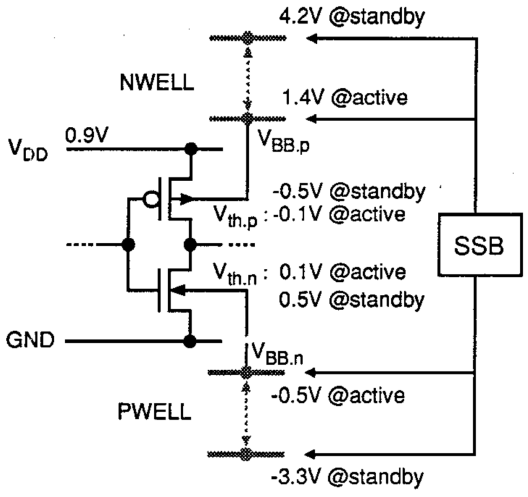


Figura 3.14: Schema VTCMOS

In *standby* invece, in riferimento alla *p-well* dell'NMOS, V_{BB} assume valori molto più bassi, comportando un aumento del grado di inversione della giunzione: ciò fa innalzare la tensione di soglia per effetto body, con conseguente diminuzione della corrente di sottosoglia del transistor. Nel caso del PMOS la tensione V_{BB} applicata è positiva e l'effetto che si ottiene è lo stesso precedentemente descritto. Riguardo SSB (*Self-Substrate Bias*), non è altro che un blocco per il controllo della tensione V_{BB}.

La VTCMOS tuttavia presenta dei limiti applicativi, dovuti soprattutto al continuo *scaling* della tecnologia [21]. Come è stato descritto, questa tecnica si avvale dell'effetto body per ottenere una variazione della tensione di soglia. Tuttavia, quando il canale del transistor inizia ad essere molto stretto, l'effetto body si "indebolisce" e l'incremento della tensione di soglia non è paragonabile a quello che si otterrebbe con tecnologie meno scalate.

3.2.5 Dynamic Threshold CMOS

A differenza della tecnica VTCMOS, la DTCMOS si propone di alterare il grado di polarizzazione dei transistor *dinamicamente*, adattandolo allo stato operativo del circuito. Per ottenere questo effetto vengono utilizzati dei transistor SOI opportunamente modificati. Infatti il *gate* ed il *bulk* sono cortocircuitati tramite un contatto metallico. La Figura 3.15 [22] illustra schematicamente la particolarità di questa tecnica.

Ne consegue che la tensione di soglia varierà con il variare della tensione dell'input.

Poniamo che il *body* del nostro transistor sia polarizzato a 0 V ($V_{bs} = 0$): la corrispondente tensione di soglia la denotiamo con V_{t0}. Poiché il *gate* ed il *bulk* sono cortocircuitati, avremmo che $V_{gs} = V_{bs}$. Supponiamo ora di aumentare la tensione V_{gs} (e quindi V_{bs}): grazie all'effetto body dovuto alla polarizzazione del *body* avremo che il valore della tensione di soglia V_t sarà minore di V_{t0}. Il grafico in Figura 3.16a a fronte [22] illustra l'andamento di V_t in funzione della tensione applicata al *gate* e quindi al *body*.

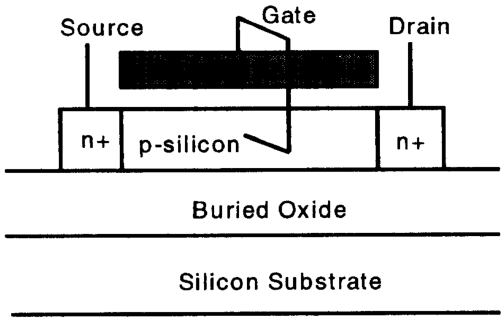


Figura 3.15: SOI NMOS con Body e Gate Cortocircuitati

Il funzionamento globale si può quindi riassumere in quanto segue:

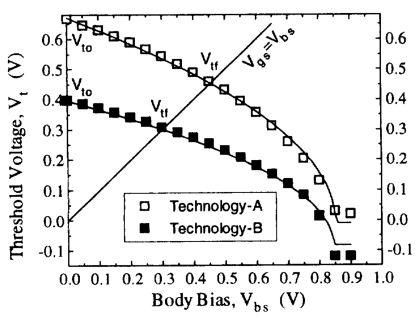
$\mathbf{V_{gs} = V_{bs} = 0 \Rightarrow V_t}$ **alta** Ciò permette di ridurre la corrente di sottosoglia quando il transistor è spento;

$\mathbf{V_{gs} = V_{bs} = V_{dd} \Rightarrow V_t}$ **bassa** In condizioni di attività, vengono migliorate le prestazioni grazie al calo della tensione di soglia.

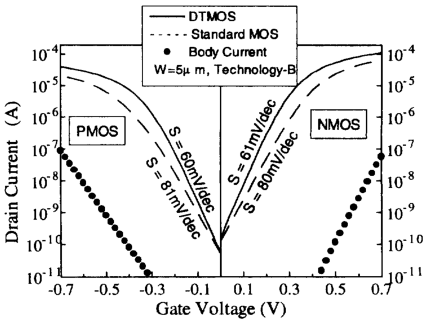
Un fatto rilevante riguarda il valore del *subthreshold slope* (vedi la nota di 2.1.1 a pagina 6): infatti il suo valore è minore ($60mV/dec$) rispetto ai classici MOSFET ($80mV/dec$) e ciò sta ad indicare una maggior efficenza di spegnimento.

Un altro vantaggio dato dalla DTCMOS è che nei singoli transistor la mobilità dei portatori è maggiore rispetto a quella che si ha nei classici MOSFET (Figura 3.16c a fronte [22]). Questo è dovuto al fatto che il campo elettrico perpendicolare al canale (dovuto alla tensione applicata al *gate*) è minore, a causa del cortocircuito fra *gate* e *bulk*.

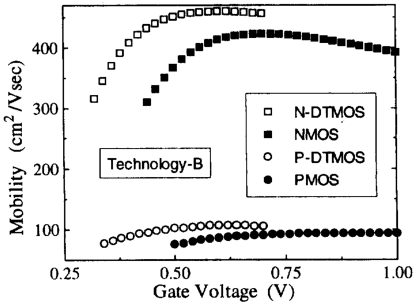
Tuttavia l'impiego della DTCMOS è limitato dalla tensione di alimentazione applicabile. Infatti il transistor può sopportare una limitata corrente



(a) Tensione di Soglia



(b) Corrente di Drain



(c) Mobilità

Figura 3.16: Tech-A: $T_{ox} = 10$nm, $N_a = 2,5 \cdot 10^{17}$ cm^{-3}. Tech-B: $T_{ox} = 6,4$nm, $N_a = 3 \cdot 10^{17}$ cm^{-3}.

attraverso di sé: questo limite determina quindi il valore della tensione di alimentazione che può essere applicato al *gate*. In Figura 3.16b [22] è illustrata (cerchietti neri) la crescita della corrente di *body* al crescere della tensione di *gate*. Tipicamente la DTCMOS non viene impiegata con tensioni di alimentazione superiori a 0,6 V, corrispondente alla soglia di accensione della giunzione *source-bulk*.

Funzionamento di un *inverter* in DTCMOS

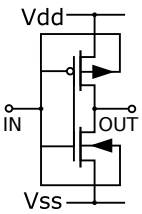


Figura 3.17

Prendiamo ora in esame il funzionamento dell'*inverter* mostrato in Figura 3.17. I casi da analizzare sono due:

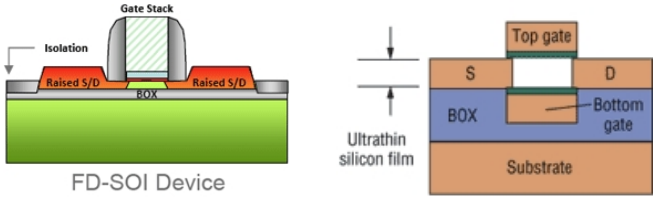
IN = $\mathbf{V_{dd}}$ In questa condizione il PMOS è spento e l'NMOS è acceso, determinando la scarica dell'OUT. Alle *well* dei due transistor viene applicata quindi una tensione positiva che fa aumentare la tensione di soglia nel PMOS, riducendo la corrente di sottosoglia, mentre la fa diminuire nell'NMOS, facilitando la scarica del nodo d'uscita.

IN $=$ $\mathbf{V_{ss}}$ Dato che il PMOS ora è acceso, e l'NMOS è spento, si ha la carica del nodo d'uscita. Le *well* ora sono a potenziale più basso rispetto al caso precedente e ciò comporta la riduzione della tensione di soglia del PMOS e l'aumento di questa dell'NMOS, con conseguenze simmetriche rispetto a quelle del caso $IN = V_{dd}$.

In definitiva quindi si può notare come questa tecnica, senza ulteriore occupazione d'area, riesca ad ottenere una riduzione sostanziale del consumo statico, senza comportare un degrado delle prestazioni.

3.2.6 Double-Gate Dynamic Threshold SOI CMOS

Come è stato sottolineato nella precedente sezione, l'uso della DTCMOS è limitato a determinati valori della tensione di alimentazione. Con la *Double-Gate Dynamic Threshold SOI CMOS* (DGDT-MOS) si riesce a sfruttare i benefici della DTCMOS senza limiti sulla tensione di alimentazione. Per far ciò si utilizza un'altra tipologia di transistor: i *Fully-Depleted SOI MOSFET* (FD SOI MOSFET).



(a) Fully Depleted SOI MOSFET

(b) FD SOI MOSFET con Front e Back Gate

Figura 3.18

Dalla Figura 3.18a [23] si può osservare che la particolarità di questi transistor risiede nel fatto di avere uno strato di silicio molto sottile al di sopra dell'isolante: ciò permette di svuotare completamente dai portatori maggioritari il *body* del transistor, ottenendo un più facile controllo sulla tensione di soglia. In Figura 3.18b [24] è mostrata la modifica tramite la quale si può gestire la tensione di soglia (questi transistor si chiamano *double-gate mosfet*, da cui il nome della tecnica) [25]. Il *front-gate* si può considerare come il classico *gate* di un MOSFET, mentre il *back-gate* (in figura *bottom gate*) serve per la polarizzazione del *body*.

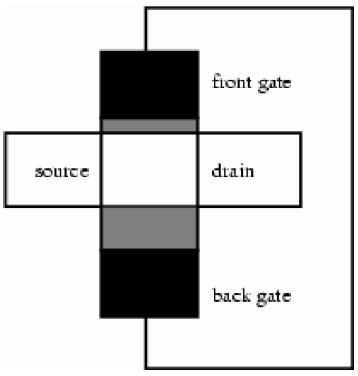


Figura 3.19: DGDT SOI MOS

In Figura 3.19 [2] è mostrato l'utilizzo che si fa di questi transistor nella DGDT SOI CMOS: il *front-gate* ed il *back-gate* sono cortocircuitati fra di loro, un po' come accade nella DTCMOS fra *gate* e *bulk*. Gli effetti che si ottengono applicando V_{dd} o V_{ss} sono uguali a quelli descritti nella sezione della DTCMOS (vedi 3.2.5 a pagina 30). Tuttavia l'impiego di questi transistor rispetto ai più semplici SOI MOSFET, permette di avere un minor consumo di potenza, poiché le capacità parassite da caricare e scaricare sono più piccole. Inoltre gli stessi sono caratterizzati da un *subthreshold slope* più piccolo (vedi la nota di 2.1.1 a pagina 6) e ciò ci dice che la corrente di sottosoglia è minore [26].

3.3 Dynamic V_{th} Designs

Le tecniche che verranno successivamente descritte sono accomunate dal fatto di ridurre il consumo di potenza del circuito quando questo è in modalità attiva. Essenzialmente si tratta di cambiare la tensione di soglia (V_{th}) dei transistor al variare del carico di lavoro del circuito: quando il carico di lavoro è alto si abbassa V_{th} per aumentare le prestazioni; viceversa, al diminuire del carico di lavoro si alza V_{th} per ridurre il consumo di potenza. Un esempio viene dato dalla Figura 3.20 [27], in cui si può notare come, all'abbassarsi della frequenza di funzionamento (e quindi del carico di lavoro), venga alzata la tensione di soglia dei transistor.

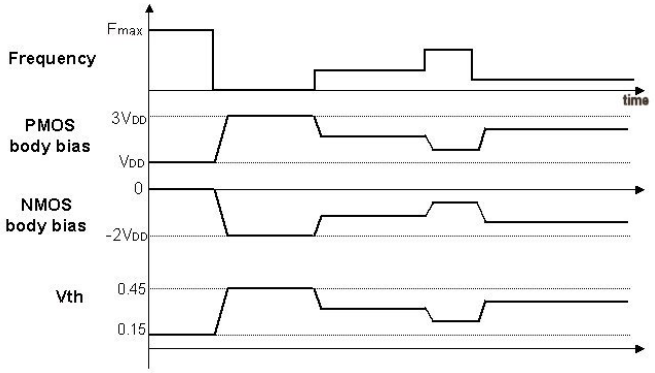


Figura 3.20: Scaling Dinamico di V_{th} tramite adattamento della Polarizzazione del Body per un Dato Profilo di Clock

Osservando il grafico in Figura 3.21 [28] si può capire l'utilità di operare nel modo sopra descritto: si può infatti notare la grande differenza che si ha fra l'uso di transistor con una fissata tensione di soglia rispetto a quelli con tensione di soglia variabile.

L'enorme *gap* è dovuto al fatto che la potenza dissipata a causa delle correnti di perdita non dipende dalla frequenza a cui opera il circuito, ma rimane costante, mentre la potenza attiva varia linearmente con la frequenza.

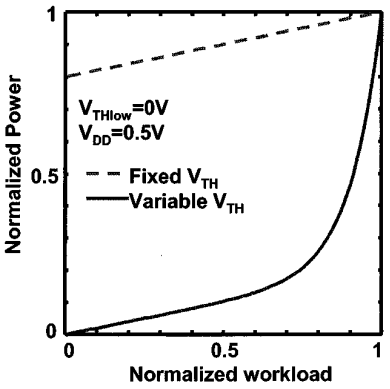


Figura 3.21: Potenza vs Carico di Lavoro Normalizzato

Per far variare la tensione di soglia di un transistor ne viene polarizzato il *bulk*, come spiegato nelle precedenti tecniche (vedi ad esempio 3.2.4 a pagina 29).

3.3.1 V_{th}-Hopping Scheme

In Figura 3.22 [28] è visibile lo schema base della *V_{th}-Hopping*. La parte a destra (*Target processor*) è il circuito che deve essere gestito, mentre il *Power control block* è, appunto, il blocco adibito all'elaborazione dei segnali per il controllo.

Il segnale CONT è un segnale elaborato via software dal processore, in base al quale vengono generati gli opportuni segnali dal *Power control block*. In [29] viene descritto uno schema a catena di retroazione con cui realizzare questa parte: anche se viene presentato per il controllo della tensione di alimentazione, è comunque possibile adeguarlo per il nostro caso.

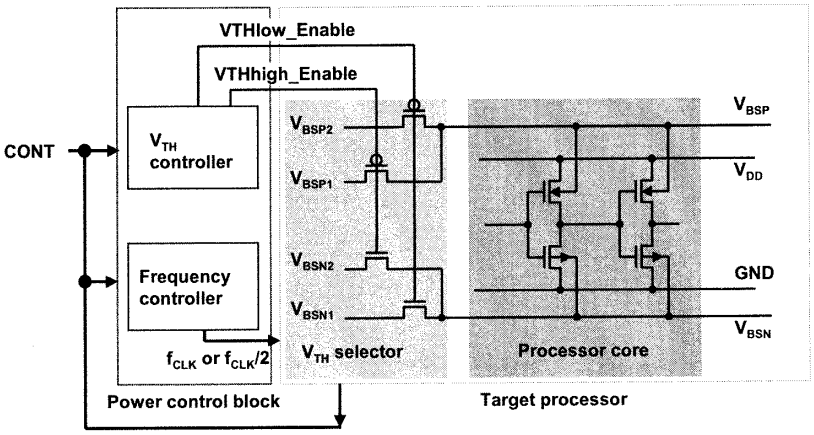


Figura 3.22: Schema V_{th}-Hopping

Per quanto riguarda il controllo della tensione di soglia, il compito è affidato al *V_{th} Controller*. Questo può generare due livelli di tensione:

- **VTHlow_Enable** farà in modo di avere una bassa tensione di soglia (V_{thLow}) nei transistor del circuito. Questo viene messo in atto per aumentare le prestazioni quando il carico di lavoro del circuito è alto;

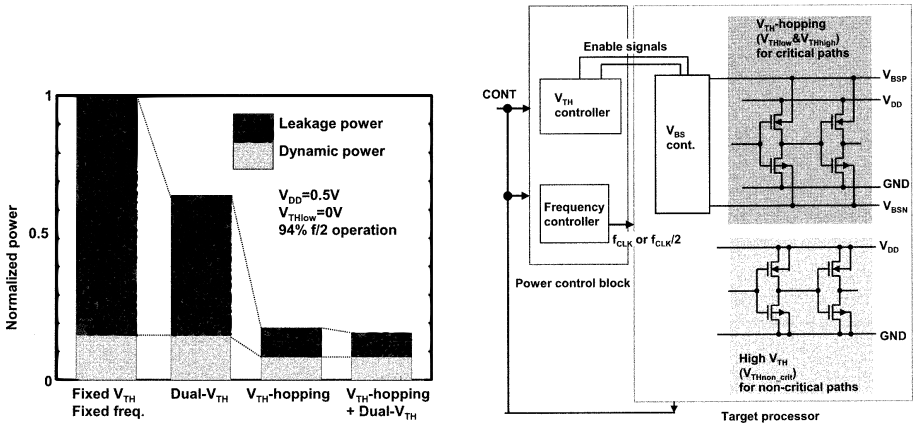
- **VTHhigh_Enable** porterà la tensione di soglia ad un livello più alto (V_{thHigh}), in condizioni di basso carico di lavoro.

Oltre a gestire la tensione di soglia, il segnale CONT serve anche a regolare la frequenza di funzionamento del circuito, tramite il *Frequency controller*. Questa operazione è comunque legata a ciò che è stato detto riguardo alla tensione di soglia: infatti nel caso in cui venga generato *VTHlow_Enable* la frequenza viene posta ad un valore massimo (f_{CLK}), mentre se il segnale elaborato è *VTHhigh_Enable*, allora la frequenza viene fissata ad un valore minore ($f_{CLK}/2$). Ovviamente V_{thLow} e V_{thHigh} devono

essere tali da permettere un buon funzionamento del circuito alle frequenze f_{CLK} e $f_{CLK}/2$, rispettivamente.

Prestazioni

In Figura 3.23a è riportato un grafico relativo al consumo di potenza (normalizzato) di circuiti che implementano delle tecniche precedentemente esposte. La simulazione è basata sull'utilizzo dell'*MPEG-4 video encoding* [28].



(a) Comparazione del Consumo di Potenza fra varie Tecniche

(b) Schema di Combinazione fra Dual-V_{th} e V_{th}-Hopping

Figura 3.23

La prima cosa che si può notare è la grande differenza di consumo data dall'utilizzo della V_{th}-*Hopping* rispetto alle altre tecniche. In secondo luogo si può notare come la combinazione (si veda Figura 3.23b) fra V_{th}-*Hopping* e *Dual Threshold CMOS* (vedi 3.2.3 a pagina 27) non apporti significativi miglioramenti rispetto all'utilizzo della sola V_{th}-*Hopping* (solo 9% in meno di consumo): nonostante la V_{th}-*Hopping* venga utilizzata solo nei cammini critici, il motivo vero di tale situazione sta nel fatto che già l'utilizzo della sola V_{th}-*Hopping* sopprime le correnti di perdita dei transitor appartenenti ai cammini critici.

3.3.2 Dynamic V_{th}-Scaling Scheme

La *Dynamic V_{th}-Scaling Scheme* (DVTS) si prefissa, come la precedente, di alterare la tensione di soglia dei transistor rispetto al carico di lavoro del circuito. Per far ciò tuttavia si utilizza un'implementazione *hardware* a differenza del *V_{th}-Hopping*. In Figura 3.24 [27] sono mostrati schematicamente i vari componenti del DVTS.

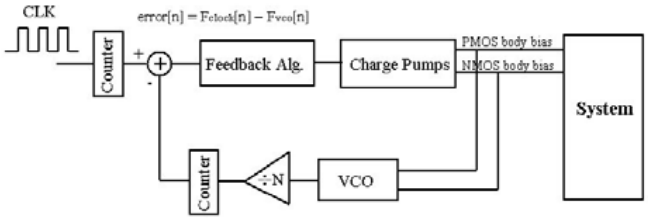


Figura 3.24: Schema Hardware del DVTS

Cominciamo con l'analizzare il blocco denominato VCO, ovvero il *Voltage Controlled Oscillator*: esso si basa essenzialmente su una catena di invertitori (vedi Figura 3.25 [27]), le cui alimentazioni sono collegate alle linee che servono a polarizzare i *body* dei MOS del circuito. Al variare delle tensioni di queste linee, varia anche la frequenza d'oscillazione di questo *ring oscillator* poiché:

$$f_{osc} \propto \frac{1}{t_p} \quad \text{e} \quad t_p \propto \frac{1}{V_{dd}}$$

in cui t_p indica il ritardo del singolo *inverter* (vedi 3.2 a pagina 17). Il segnale elaborato da questa unità è proporzionale al ritardo del percorso critico che in quel momento viene "usato" nel circuito.

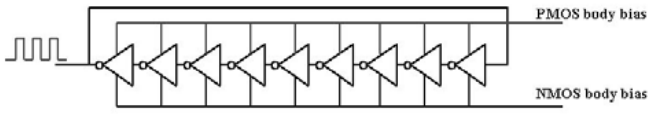


Figura 3.25: Catena di Inverter del VCO

Il segnale elaborato da questa unità (una frequenza) viene poi confrontato con un segnale di *clock* di riferimento (in alto a sinistra in Figura 3.24). La differenza fra questi due segnali sarà utilizzata da un algoritmo di *feedback* che gestisce una pompa di carica (*charge pump*) con cui andare a modificare le tensioni di polarizzazione dei transistor.

3.4 Supply Voltage Scaling

Come è stato sottolineato più volte nel corso di questo elaborato, la maniera più efficacie per ridurre il consumo di potenza dinamico è quello di diminuire la tensione di alimentazione, dato il suo legame quadratico con la potenza (vedi 3.2 a pagina 17). Le tecniche che stiamo andando a considerare, prendono spunto proprio da quanto appena detto e si differenziano per la modalità di riduzione della tensione di alimentazione: in una si applicherà una riduzione *statica*, che non varia al variare di determinati parametri; nell'altra invece la riduzione sarà *dinamica*, cioè cambierà rispetto, ad esempio, al carico di lavoro del circuito.

Tali tecniche tuttavia sono sottoposte a dei limiti nei possibili valori applicabili (vedi Figura 1.2 a pagina 2 e relativa descrizione).

Static Supply Voltage Scaling

L'idea principale in questo caso è quella di individuare, nel circuito, dei percorsi critici e dei percorsi non critici, a cui verranno applicate due diverse tensioni di alimentazione. In Figura 3.26 [2] viene espresso schematicamente questo concetto.

La definizione di percorso critico è la stessa che è stata data in 3.2.3 a pagina 27. Sapendo che il ritardo di una porta dipende in maniera inversa dal valore della tensione di alimentazione (vedi 3.2 a pagina 17), si capisce il motivo per cui venga assegnata ad un percorso critico una V_{dd} alta e, viceversa, ad un percorso non critico venga attribuito un basso valore di V_{dd}. Ciò non va ad intaccare le prestazioni del sistema complessivo, poiché la velocità richiesta nei percorsi non critici è più bassa di quella prevista dei percorsi critici.

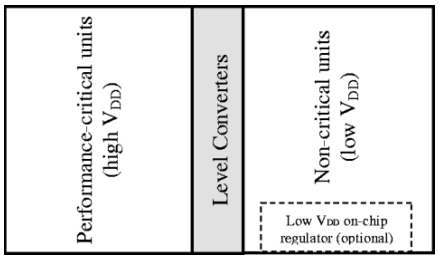


Figura 3.26: Static Supply Voltage Scaling

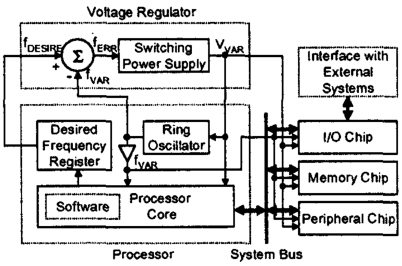
La funzione del *level converter* riportato in Figura 3.26 riguarda l'adeguamento dei segnali fra una zona e l'altra: se infatti degli input della zona ad alta tensione di alimentazione provengono da quella a bassa tensione di alimentazione, è necessario convertirli al valore adeguato.

Dynamic Supply Voltage Scaling

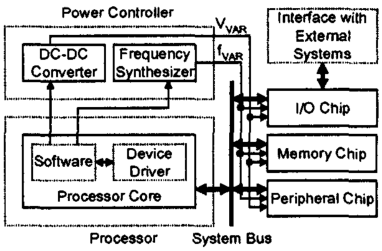
A differenza della tecnica precedente, la *Dynamic Supply Voltage Scaling* si prefigge come obiettivo quello di adeguare dinamicamente il valore della

tensione di alimentazione rispetto al carico del circuito. Un modo per far ciò è rappresentato dall'architettura *DVS*, che sta per *Dynamic Voltage Scaling*, di cui è riportata un'implementazione in Figura 3.27a [29]: la tensione di alimentazione è controllata tramite un anello di retroazione che utilizza un oscillatore ad anello (*ring oscillator*) per simulare il percorso critico (un po' come descritto in 3.3.2 a pagina 37). Tuttavia questa modalità presenta dei problemi [29]:

- La relazione tensione di alimentazione-frequenza non è la stessa per ogni circuito, dato che il processo di fabbricazione è diverso per ogni sistema: ciò purtroppo non viene tenuto in considerazione da questa implementazione;
- Poiché il *ring oscillator* deve essere inserito nel circuito, se abbiamo dei sistemi "preconfezionati" (come ad esempio i processori commerciali per gli sviluppatori) è impossibile utilizzare tale architettura;
- Se siamo in presenza di un sistema multi-processore, l'idea sarebbe quella di poter controllare la tensione di alimentazione in maniera indipendente per ogni processore, cosa che non è possibile con questa modalità.



(a) Architettura DVS Convenzionale



(b) Architettura DVS Proposta in [29]

Figura 3.27

Per risolvere questi problemi, sempre in [29] viene proposta una diversa architettura che viene riportata in Figura 3.27b. Le modifiche apportate tengono conto di vari fattori, fra cui:

- La tensione di alimentazione viene controllata con *feedback* tramite software;
- La determinazione della tensione di alimentazione da applicare avviene basandosi sulla relazione tensione-frequenza di ogni chip indipendentemente;

- La frequenza di clock del sistema può assumere solo valori discreti ($f_{clk}, f_{clk}/2, f_{clk}/3 \ldots$) per evitare problemi di interfacciamento fra il *controller* e il processore. f_{clk} è la frequenza del *clock* di riferimento.

Il sistema presenta poi due tabelle di riferimento, ricavate dalle proprietà fisiche del circuito: una riguarda le relazioni tensione-frequenza, l'altra i ritardi di transizione nel cambio di frequenza e di tensione di alimentazione.

Questa architettura risulta più flessibile della precedente, in quanto può essere implementata sia come parte integrante del circuito, sia come componente esterno.

3.5 Clock Gating

Nei sistemi sincroni è di fondamentale importanza la presenza del segnale di *clock*: in questi sistemi infatti le commutazioni possono avvenire in corrispondenza di uno dei due fronti di *clock*. Di conseguenza, la rete che porta il *clock* nei vari punti del circuito è molto sviluppata e negli ultimi anni si è rivelata essere una delle principali fonti di consumo di potenza. Inoltre per evitare che il *clock* arrivi ai diversi componenti del circuito in tempi diversi[3], a causa ad esempio di differenti lunghezze delle piste o per la presenza di capacità parassite, si introducono dei *buffer* per poter gestire meglio il ritardo sulle diverse linee. Ciò non fa altro che aggiungere carico alla rete del *clock* e quindi aumenta ulteriormente il consumo di potenza.

Per risolvere questo problema e ridurre la potenza usata dal *clock*, si fraziona la rete e si "creano" altri *clock*, derivati da quello principale (chiamato *master clock*). In questo modo si diminuisce il carico della rete del *master clock* e con esso anche il numero di *buffer*. Inoltre si riduce anche la potenza complessiva consumata, in quanto:

- La frequenza dei nuovi *clock* può essere adeguata al carico di lavoro della parte di circuito in cui si trovano;
- I *flip-flop* connessi ai *clock* derivati non commutano nei cicli di inattività.

Ovviamente questi *clock* non sono totalmente indipendenti dal *clock master*: per derivare in maniera corretta i *clock* esistono vari metodi, alcuni dei quali esposti in [30].

[3] Effetto noto come *clock skew*.

3.6 Voltage Island

Con l'avanzare della tecnologia e il continuo *scaling* dei componenti fondamentali, l'implementazione di sistemi complessi è molto cambiata: si è passati infatti da *design* che prevedevano la costruzione di vari blocchi funzionali "indipendenti" fra di loro, a *design* in cui l'intero sistema viene costruito in un unico chip di silicio (*System-on-a-Chip*). La gestione, ad esempio, della tensione di alimentazione e della frequenza di funzionamento del circuito non è quindi più particolareggiata per ogni singolo blocco, ma risulta ora unica. La tecnica denominata *Voltage Island* si propone di frazionare il circuito in gruppi (*island*) di *core* che operano alla stessa tensione di alimentazione, in modo da ottimizzare il consumo di potenza modificando le singole tensioni delle varie sezioni.

In Figura 3.28 [31] si ritrae schematicamente la suddivisione di un chip, con a fianco i valori di tensione permessi nelle varie sezioni. Nell'esempio proposto, le *voltage island* non coincidono con le sezioni mostrate: infatti si può notare come c_1, c_3 e c_4 abbiano le stesse tensioni di funzionamento permesse e quindi possono costituire una *voltage island* unica. Tuttavia, per raggruppare insieme diversi *core* bisogna tener conto di eventuali problemi di *design* come la sincronizzazione delle varie aree e la congestione delle interconnessioni [31].

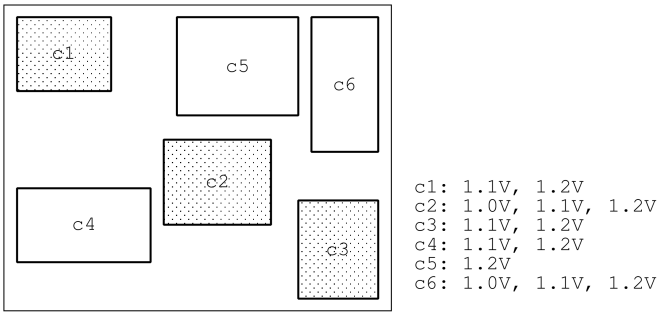
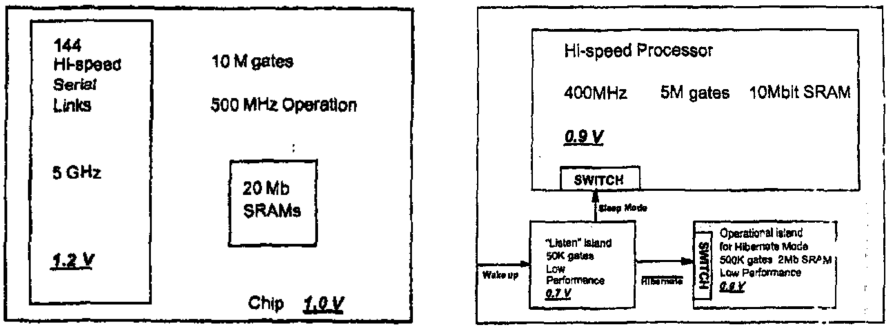


Figura 3.28: Voltage Island

Esistono varie situazioni in cui la *Voltage Island* può essere applicata: prendiamo ad esempio gli schemi presentati in Figura 3.29 nella pagina successiva[32].

In Figura 3.29b a fronte viene presentato un circuito sempre (o quasi) attivo: in tal caso l'applicazione delle *Voltage Island* permetterà di ridurre il consumo di potenza attiva. Nella parte di circuito caratterizzato da alta criticità, come ad esempio un processore, per ottenere alte prestazioni si applica un'alta tensione di alimentazione; viceversa nelle parti meno critiche, come la memoria, la tensione può essere tranquillamente più bassa.



(a) Voltage Island in un Dispositivo Critico (b) Voltage Island in un Power-Sequencing

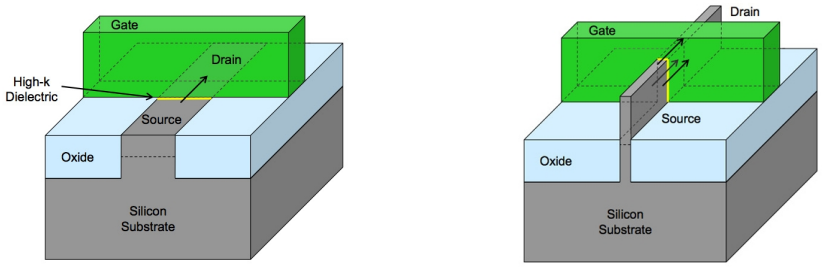
Figura 3.29

In Figura 3.29a viene invece raffigurato un circuito in cui risulta importante il risparmio di energia, come può essere un sistema alimentato a batterie. In apparati del genere, molte funzionalità non sono attive per la maggior parte del tempo e, di conseguenza, si possono avere alte correnti di perdita se l'alimentazione resta attiva anche in quelle porzioni del circuito. Tramite la *Voltage Island* si riesce a disattivare l'alimentazione delle porzioni che sono in *standby* e ridurre al minimo la potenza dovuta alle correnti di perdita.

3.7 FinFET

Dalla loro creazione nel 1959, i MOSFET sono stati sempre più rimpiccioliti ma hanno conservato comunque la loro forma originaria, ovvero quella planare, mostrata in Figura 3.30a. Tuttavia le dimensioni a cui la tecnologia attuale opera portano ad avere una sempre maggiore incidenza dei cosidetti *effetti di canale corto* (SCEs), come l'*hot carrier injection* e la saturazione di velocità dei portatori, ed anche una sempre più alta rilevanza della corrente di sottosoglia nel consumo totale. Ciò è dovuto al fatto che, dal punto di vista della grandezza, le zone di svuotamento delle giunzioni del *drain* e del *source* diventano comparabili con il canale: la variazione della lunghezza di quest'ultimo, di conseguenza, risulta non più trascurabile, comportando l'amplificazione di effetti parassiti, che nei dispositivi a canale lungo erano trascurabili.

I problemi però non si esauriscono qui: infatti per limitare l'intensità degli SCEs si è proceduto ad assottigliare sempre più lo strato di ossido sopra al *gate*, comportando la possibilità per gli elettroni e le lacune di passare attraverso questo (vedi 2.2 a pagina 9). Come se non bastasse, la piccola distanza fra *drain* e *source* fa sì che degli elettroni caratterizzati da un'energia più alta della media, possano uscire dalla parte inferiore del canale, provocando ulteriore corrente di perdita e dissipazione di calore.



(a) MOSFET Planare (b) FinFET

Figura 3.30

Per poter risolvere i problemi sopra elencati e continuare a ridurre le dimensioni dei transistor, nel 1999 un team della *University of California, Berkeley* ha descritto per la prima volta una nuova struttura di transistor, sempre ad effetto di campo, denominato *Finfet* (*Fin-shaped Field Effect Transistor*). In Figura 3.30b si può osservare la particolarità di questa nuova struttura: i vari componenti del transistor (*drain*, *source*, *gate* e *bulk*) non si espandono più solo lungo un piano, ma anche in altezza. Il canale che collega *drain* e *source* si sviluppa ora in altezza, mantenendo comunque uno spessore molto limitato.

I vantaggi sono molteplici, dovuti soprattutto al fatto che ora il *gate* avvolge il canale da tre lati, mentre prima lo copriva solo lungo una superficie. Grazie a questo fatto si ha:

- Riduzione degli elettroni che riescono a lasciare il canale, poiché, energicamente parlando, vengono controllati meglio dalla tensione al *gate* e quindi saranno in un numero minore quelli ad un'energia più alta, rispetto al caso del MOSFET planare. Inoltre lo spazio in cui tali elettroni possono "fuggire" è ridotto dalla presenza del *gate* lungo i tre lati;
- Riduzione della corrente di sottosoglia, a causa del fatto che la tensione applicata al *gate* può agire più efficacemente, rispetto a prima, nel controllo degli elettroni liberi nel canale. Di conseguenza la tensione di soglia può essere abbassata senza provocare una maggiore corrente di sottosoglia;
- Riduzione degli effetti di canale corto.

L'utilizzo di questi transistor in prodotti commerciali tuttavia, prevede cambiamenti anche molto importanti nell'industria dei semiconduttori, a causa soprattutto del cambio della struttura e quindi della necessità di riuscire a sviluppare i transistor anche in altezza e non più solo in larghezza e lunghezza.

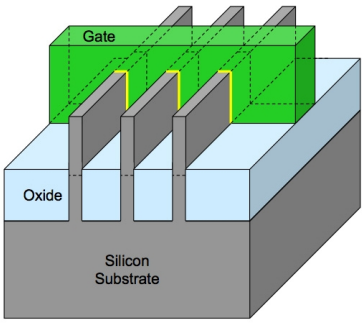


Figura 3.31: Intel Tri-Gate

Intel ha annunciato in maggio 2011 di aver avviato la produzione di processori a 22 nm basati su transistor a struttura tridimensionale, chiamati *Tri-gate* [33] e mostrati in Figura 3.31. Dai dati riportati sempre in [33], si evince come, rispetto a processori a 32 nm con transistor planari, vi sia, in condizioni di bassa tensione di alimentazione, un aumento delle prestazioni del 37% ed un dimezzamento della potenza consumata, a parità di *performance*.

4. Conclusioni

Abbiamo analizzato molte tecniche in questo lavoro e tutte apportano benefici al consumo di potenza. Tuttavia non sono tutte equivalenti: la scelta di una tecnica piuttosto che un'altra risiede ad esempio nella possibilità di intervenire ad un determinato livello del *design* del circuito piuttosto che un altro. Basti pensare alla *Transistor Stacks* ed alla DTCMOS: la prima interviene a livello puramente circuitale, mentre la seconda si applica al singolo dispositivo. Un'altra causa di discriminazione può riguardare i limiti applicativi: ad esempio la VTCMOS, come abbiamo detto, non può essere utilizzata in circuiti con transistor troppo piccoli, mentre l'impiego della DTCMOS è limitato dal valore della tensione di alimentazione. Un ulteriore fattore limitante per certe tecniche può essere rappresentato dal dover implementare ulteriori blocchi logici rispetto al circuito in considerazione: si veda ad esempio quanto scritto riguardo le due architetture DVS presentate in 3.4. Ogni tecnica presenta pregi e difetti: dipende quindi dal progettista sfruttare quella più opportuna in relazione al circuito preso in esame.

L'avvento dei nuovi transistor tridimensionali può limitare certi problemi (come l'eccessiva corrente di sottosoglia), ma con l'inarrestabile integrazione e con l'aumento della richiesta di dispositivi *mobile*, la ricerca di tecniche per la riduzione del consumo di potenza procederà verso nuovi risultati.

Bibliografia

[1] Online: accessed 14 September 2011. Intel Corporation. 2011. URL: http://www.intel.com/content/dam/staging/image/Kim/HTML%20Detail%20Pages/moores_Infographic_2-1.jpg.

[2] Kaushik Roy, Saibal Mukhopadhyay e Hamid Mahmoodi-Meimand. "*Leakage Current Mechanisms and Leakage Reduction Techniques in Deep-Submicrometer CMOS Circuits*". In: *Proceedings of the IEEE* (2003), pp. 305–327.

[3] Walid M. Elgharbawy e Magdy A. Bayoumi. "*Leakage Sources and Possible Solutions in Nanometer CMOS Technologies*". In: *IEEE Circuit and Systems Magazine* Fourth Quarter (2005), pp. 6–17.

[4] Andrea Cester. Online: accessed 15 September 2011. Department of Information Engineering, University of Padua. 2009. URL: www.dei.unipd.it/~cester/download/corsi/scaling-consumo.pdf.

[5] *International Technology Roadmap for Semiconductors*. Online: accessed 12 August 2011. International SEMATECH. 2009. URL: http://public.itrs.net/.

[6] Ndubuisi Ekekwe e Ralph Etienne-Cummings. "*Power Dissipation Sources and Possible Control Techniques in Ultra Deep Submicron CMOS Technologies*". In: *Microelectronics Journal* 37 (2006), pp. 851–860.

[7] Jan M. Rabaey, Anantha Chandrakasan e Borivoje Nikolic´. Online: accessed 03 September 2011. 2003. URL: http://bwrc.eecs.berkeley.edu/IcBook/Slides/chapter5.ppt.

[8] Jan M. Rabaey, Anantha Chandrakasan e Borivoje Nikolic´. *Circuiti Integrati Digitali. L'ottica del Progettista*. Pearson Education Italia, 2005.

[9] *High-k and Metal Gate Research*. Online: accessed 25 Semptember 2011. Intel Corporation. 2007. URL: http://www.intel.com/technology/silicon/high-k.htm.

[10] Naoki Kotani et al. "*An Impact of GIDL Off Leakage on Low-Power Sub-0.2µm SOI CMOS Applications*". In: *IEEE International SOI Conference* (2000), pp. 90–91.

[11] Jonathan P. Halter e Farid N. Najm. "*A Gate-Level Leakage Power Reduction Method for Ultra-Low-Power CMOS Circuits*". In: *Proceedings IEEE Custom Integrated Circuits Conference* (1997), pp. 475–478.

[12] Mark C. Johnson, Dinesh Somasekhar e Kaushik Roy. "*Leakage Control With Efficient Use of Transistor Stacks in Single Threshold CMOS*". In: *ACM/IEEE Design Automation Conference* (1999), pp. 442–445.

[13] Yibin Ye, Shekhar Borkar e Vivek De. "*A New Technique for Standby Leakage Reduction in High-Performance Circuits*". In: *Symposium on VLSI Circuits Digest of Technical Papers* (1998), pp. 40–41.

[14] David Duarte et al. "*Evaluating Run-Time Techniques for Leakage Power Reduction*". In: *Proceedings of the 15th International Conference on VLSI Design* (2002), pp. 31–38.

[15] Shin'ichiro Mutoh et al. "*1-V Power Supply High-Speed Digital Circuit Technology with Multithreshold-Voltage CMOS*". In: *IEEE Journal of Solid-State Circuits* (1995), pp. 847–854.

[16] Online: accessed 10 September 2011. Seiko Instruments Inc. 2011. URL: `http://www.sii.co.jp/info/eg/soi3.html`.

[17] James Kao, Anantha Chandrakasan e Dimitri Antoniadis. "*Transistor Sizing Issues and Tool For Multi-Threshold CMOS Technology*". In: *Design Automation Conference, Proceedings of the 34th ACM/IEEE* (1997), pp. 409–414.

[18] Hiroshi Kawaguchi, Koichi Nose e Takayasu Sakurai. "*A Super Cut-Off CMOS (SCCMOS) Scheme for 0.5-V Supply Voltage with Picoampere Stand-By Current*". In: *IEEE Journal of Solid-State Circuits* (2000), pp. 1498–1501.

[19] Liqiong Wei et al. "*Design and Optimization of Dual-Threshold Circuits for Low-Voltage Low-Power Applications*". In: *IEEE Transactions on Very Large Scale Integration Systems* (1999), pp. 16–24.

[20] Tadahiro Kuroda et al. "*A 0.9V 150MHz 10mW 4 mm^2 2-D Discrete Cosine Transform Core Processor with Variable-Threshold-Voltage Scheme*". In: *IEEE International Solid-State Circuits Conference* (1996), pp. 166–167, 437.

[21] A. Keshavarzi et al. "*Effectiveness of Reverse Body Bias for Leakage Control in Scaled Dual Vt CMOS ICs*". In: *International Symposium on Low Power Electronics and Design* (2001), pp. 207–212.

[22] Fariborz Assaderaghi et al. "*A Dynamic Threshold Voltage MOSFET (DTCMOS) for Ultra-Low Voltage Operation*". In: *IEEE International Technical Digest Electron Devices Meeting* (1994), pp. 809–812.

[23] Online: accessed 12 September 2011. UBM Electronics. 2011. URL: http://www.edn.com/article/512696-Fully_depleted_SOI_shows_its_stuff_in_CPU_design.php.

[24] Online: accessed 12 September 2011. PennWell Corporation. 2011. URL: http://www.electroiq.com/articles/sst/print/volume-47/issue-8/features/device-engineering/an-analytical-look-at-vertical-transistor-structures.html.

[25] Liqiong Wei, Zhanping Chen e Kaushik Roy. "*Double Gate Dynamic Threshold Voltage (DGDT) SOI MOSFETs for Low Power High Performance Designs*". In: *Proceedings of IEEE SOI Conference* (1997), pp. 82–83.

[26] Andrew Marshall e Sreedhar Natarajan. "*PD-SOI and FD-SOI: a Comparison of Circuit Performance*". In: *9th International Conference on Electronics, Circuits and Systems* (2002), pp. 25–28.

[27] Chris H. Kim e Kaushik Roy. "*Dynamic V_TH Scaling Scheme for Active Leakage Power Reduction*". In: *IEEE Proceedings of Design, Automation and Test* (2002), pp. 163–167.

[28] Koichi Nose et al. "*V_{TH}-Hopping Scheme to Reduce Subthreshold Leakage for Low-Power Processors*". In: *IEEE Journal of Solid-State Circuits* (2002), pp. 413–419.

[29] Seongsoo Lee e Takayasu Sakurai. "*Run-time Voltage Hopping for Low-power Real-time Systems*". In: *IEEE/ACM Proceedings Design Automation Conference* (2000), pp. 806–809.

[30] Quing Wu, Massoud Pedram e Xunwei Wu. "*Clock-Gating and Its Application to Low Power Design of Sequential Circuits*". In: *IEEE Transactions on Circuits And Systems I, Reg. Papers* (2000), pp. 415–420.

[31] Jingcao Hu et al. "*Architecting Voltage Islands in Core-based System-on-a-Chip Designs*". In: *Proceedings of International Symposium on Low Power Electronics and Design* (2004), pp. 180–185.

[32] David E. Leackey, Paul S. Zuchowski e Thomas R. Bednar. "*Managing Power and Performance for System-on-Chip Designs using Voltage Island*". In: *IEEE/ACM Conference on Computer Aided Design* (2002), pp. 195–202.

[33] *Intel Reinvents Transistors Using New 3-D Structure*. Online: accessed 15 September 2011. Intel Corporation. 2011. URL: http://newsroom.intel.com/community/intel_newsroom/blog/2011/05/04/intel-reinvents-transistors-using-new-3-d-structure.

[34] Rachel Courtland. *The Origins of Intel's New Transistor, and Its Future*. Online: accessed 15 Semptember 2011. IEEE. 2011. URL: http://spectrum.ieee.org/semiconductors/design/the-origins-of-intels-new-transistor-and-its-future.