POLITECNICO DI MILANO

Facoltà di Ingegneria

Corso di Laurea Magistrale in
Ingegneria Elettronica

Simulazioni Monte Carlo del Leakage Multitrappola e Impatto sull'Affidabilità di Dispositivi Scalati

Relatore: Prof. Alessandro Sottocornola Spinelli

Tesi di Laurea Magistrale di:
Pietro Zambon
Matricola: 721197

Anno Accademico 2009-2010

Indice

Elenco delle Figure

Elenco delle Tabelle

Riassunto

Capitolo 1 Introduzione alla tecnologia delle memorie FLASH, con una presentazione dei due principali tipi di architetture: NAND e NOR, di come avviene la programmazione/cancellazione e la lettura del dato memorizzato. Breve digressione sullo *scaling* degli ossidi di *gate* in transistori MOS ultra–scalati, presentando l'innovativa tecnologia *High-k*;

Capitolo 2 Studio dei modelli di conduzione di corrente attraverso i *layers* di *gate*, in particolare quelli di *tunneling* tradizionale di tipo diretto e di Fowler-Nordheim, in approssimazione semi–classica WKB, quello monodimensionale 1TAT (*Trap Assisted Tunneling*), la sua evoluzione al 2TAT analitico e la comparazione dei risultati numerici di quest'ultimo con dati sperimentali sulla corrente di *gate* di un transistore MOS–HK. Estensione del modello dal monodimensionale al tridimensionale approssimato del modello TAT ad un numero N arbitrario di difetti, intesi ora localizzati spazialmente anziché considerarne la densità. Verifica della coerenza del nuovo modello approssimato col precedente tramite simulazioni sullo stesso transistore MOS–HK;

Capitolo 3 Esposizione formale dei principali metodi di generazione di variabili *random* non uniformi utilizzati nel corso delle simulazioni. Presentazione del concetto di simulazione Monte Carlo standard e dei due più importanti metodi di *enhancement* statistico: *Importance Samplig* e *Splitting Tecnique*. Applicazione di tali metodi per la determinazione della distribuzione statistica della corrente di *leakage* di un *array* di memorie FLASH sottoposto a cicli di programmazione/cancellazione e confronto con i relativi dati sperimentali;

Capitolo 4 Studio sull'affidabilità in ritenzione di un'*array* di memorie FLASH di tipo NOR multilivello, in particolare al variare del criterio di fallimento, dello spessore dell'ossido di *gate* e della densità di trappole all'interno di esso, in considerazione di una previsione di *scaling* del suo spessore al di sotto degli attuali 9–10nm. Esposizione ed applicazione del metodo di ritenzione in accelerazione di tensione sul medesimo *array*.

A mio padre e a mia madre.

Introduzione Generale

I continui miglioramenti mostrati nel corso degli anni da parte delle industrie di semiconduttore nel campo delle tecnologie per la lavorazione del silicio hanno portato ad una continua miniaturizzazione dei dispositivi elettronici. Questi sviluppi tecnologici impattano positivamente sia sulle prestazioni sia sui costi di produzione del dispositivo. In particolare, nel segmento delle memorie, entrambi gli aspetti giocano un ruolo fondamentale per lanciare sul mercato un prodotto di largo successo. Negli ultimi anni, il settore delle memorie non-volatili è diventato uno dei settori che garantisce fatturati maggiori per le industrie microelettroniche. Tra queste ultime, ad occupare un posto di primissimo piano sono le memorie FLASH, delle quali la presente trattazione approfondirà alcune problematiche.

Allo stato dell'arte, si è giunti ad un livello di *scaling* tale per cui le dimensioni caratteristiche sono di circa $25nm$ per quanto riguarda le memorie FLASH con architettura NAND, utilizzate in tutte le applicazioni che richiedono l'immagazzinamento di dati ad alta densità, mentre sono di circa $45nm$ per quanto riguarda le memorie FLASH ad architettura NOR, i cui requisiti di densità sono meno stringenti per via del fatto che devono, tipicamente, memorizzare codice e non dati, con la controparte di poter garantire un'affidabilità molto più elevata: nel codice memorizzato non sono consentiti errori. La richiesta di sempre maggiore affidabilità è però in contrasto con la miniaturizzazione dei dispositivi poiché questa comporta anche la riduzione dello spessore dell'ossido di *gate*, che in questo modo risulta ancora più "trasparente" agli elettroni che lo attraversano per effetto *tunnel*, fenomeno ovviamente indesiderato e deleterio per il funzionamento della cella. La situazione è ancor peggiore a seguito della ciclatura (programmazione/cancellazione) perché causa, all'interno dell'ossido, una generazione di difetti reticolari che fungono da "ponte" per il passaggio di elettroni, incrementando, a nostro svantaggio, l'indesiderata perdita di carica – fenomeno chiamato *Stress Induced Leakage Current* (SILC). Da qui, dunque, l'importanza di sviluppare opportuni modelli di affidabilità per le celle di memoria, così da trarne preziose indicazioni sulle direzioni da seguire all'aumentare dello *scaling*, in particolare dell'ossido di *gate*, e per l'impatto che la densità di difetti reticolari, anche detti stati trappola, ha sulla probabilità di fallimento dell'*array*.

Capitolo 1

Introduzione alle memorie FLASH e prospettive di scaling per la tecnologia CMOS

A good gulp of hot whiskey at bedtime.
It's not very scientific but it helps.

(Un bel sorso di whisky caldo
prima di andare a dormire.
Non è molto scientifico, ma aiuta.)

Alexander Fleming

Nel presente capitolo verrà effettuata una panoramica introduttiva sulla tecnologia delle memorie FLASH, indicandone le caratteristiche salienti dal punto di vista elettronico e le prospettive per il futuro. Infine si avrà un breve compendio sulla tecnologia High–k*, mostrandone i principali vantaggi e svantaggi.*

§1.1 Le memorie FLASH

Prima di entrare nei dettagli della discussione sulla tecnologia delle memorie FLASH, è utile concedere una breve introduzione sul funzionamento del transistore MOSFET, sui principi del quale la suddetta tecnologia si basa.

Il *Metal–Oxide–Semiconductor Field–Effect–Transistor* (MOSFET) è il dispositivo elettronico più importante per quanto riguarda circuiti integrati di larga scala, come microprocessori e memorie a semiconduttore. La sua struttura è quella di un dispositivo a quattro terminali e consiste di un substrato di materiale semiconduttore drogato di tipo p o n, chiamato *bulk*, entro il quale sono formate due regioni di tipo n^- o p^+, rispettivamente, che sono chiamate *source* e *drain*. La parte superiore di semiconduttore, tra il *source* ed il *drain*, è coperta da uno strato isolante, tipicamente diossido di silicio (SiO_2), caratterizzato dalla sua permettività dielettrica ϵ_{ox}. Un contatto di metallo, chiamato *gate*, è successivamente deposto sopra lo strato di ossido. Il *gate* può essere di alluminio o di altri tipi di metallo, anche se nella maggior parte dei casi si tratta di silicio policristallino altamente conduttivo. La più importante caratteristica del transistore MOSFET è la tensione di soglia, *i.e.* il voltaggio da applicare al *gate* per raggiungere la condizione di inversione. La condizione di inversione è definita come quella condizione per la quale il potenziale superficiale Φ_s è uguale a $2\Phi_{Fp}$, per un semiconduttore di tipo p, o uguale a $2\Phi_{Fn}$, per un conduttore di tipo n; dove Φ_F è la distanza, in energia, tra il livello di Fermi intrinseco e quello estrinseco. Osservando la situazione appena descritta da un punto di vista fisico, in condizione di soglia la concentrazione di portatori minoritari (elettroni per un semiconduttore di tipo p, lacune per uno di tipo n) raggiunge la stessa concentrazione dei portatori maggioritari (elettroni per un semiconduttore di tipo n, lacune per uno di tipo p).

Altre importanti caratteristiche del MOSFET sono lo spessore dell'ossido t_{ox}, che è la distanza tra la superficie del substrato e il contatto di *gate*, e la lunghezza di canale L, che è la distanza tra i contatti di *source* e *drain*. La rappresentazione schematica di un dispositivo MOSFET è mostrata in Fig. 1.1, mentre i relativi simboli circuitali sono mostrati in Fig. 1.2.

Negli equipaggiamenti elettronici esistono molti differenti tipi di dispositivi MOSFET. Tuttavia, per quanto ci riguarda, ci riferiremo sempre alla struttura sopra descritta, in particolare a quella di un n–MOSFET.

La cella di memoria FLASH è essenzialmente un transistore MOSFET in cui è però presente un ulteriore strato conduttivo all'interno dello strato isolante, tra substrato e *gate*. Questo elettrodo è chiamato *floating gate*. Una sezione schematica lungo la direzione di canale di un *floating gate* transistor è mostrata in figura Fig. 1.3.

Questo elettrodo è completamente isolato elettricamente – e per questo è chiamato *floating gate* (FG) – sia dal *gate* che dal *source* che dal *drain* che

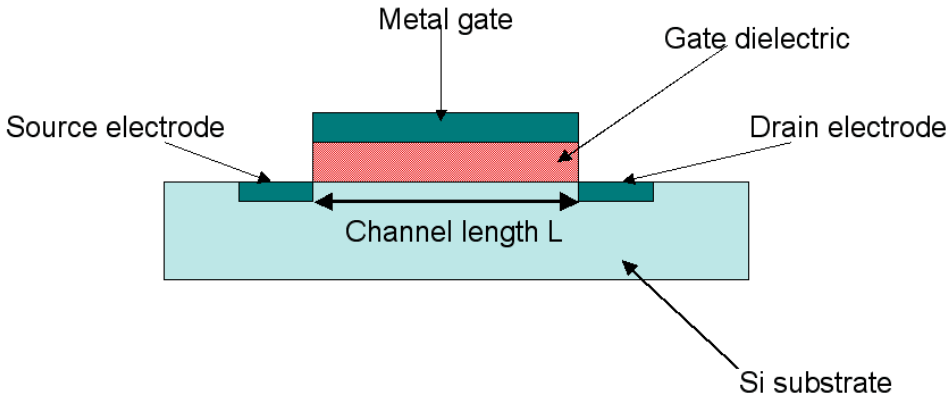


Figura 1.1: *Rappresentazione della sezione longitudinale di un MOSFET. I contatti e alcune importanti caratteristiche, come la lunghezza di canale, sono indicati.*

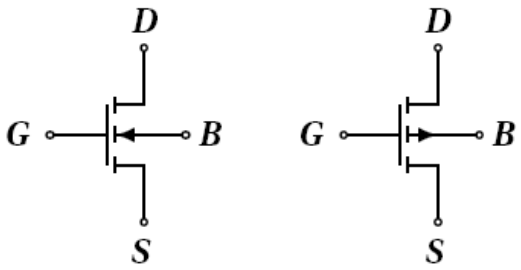


Figura 1.2: *Simboli ciruitali per i due tipi di transistor MOSFET. Nella figura di sinistra è rappresentato un n–MOSFET, mentre a destra un p–MOSFET.*

dal *bulk*, attraverso uno strato di ossido di alta qualità cresciuto termicamente su un substrato di silicio con direzione cristallografica $\langle 100 \rangle$. Questo strato isolante rappresenta l'elemento più critico per quanto concerne l'affidabilità della memoria. Infatti, come sarà chiaro più avanti, l'isolante deve garantire il miglior isolamento possibile per il FG durante la fase di ritenzione del dato, *i.e.* mantenere la carica all'interno del FG stesso, ma allo stesso tempo deve consentire uno efficace scambio di carica dal substrato al FG e viceversa in fase di programmazione/cancellazione. Questo trasferimento di carica deve avvenire in tempi ragionevolmente brevi e con l'applicazione di basi voltaggi di controllo.

Ovviamente, queste richieste sono in contrasto tra loro ed è necessario raggiungere un *trade–off* per lo spessore dell'ossido, di qualità microscopica, come parametro di *scaling* fondamentale.

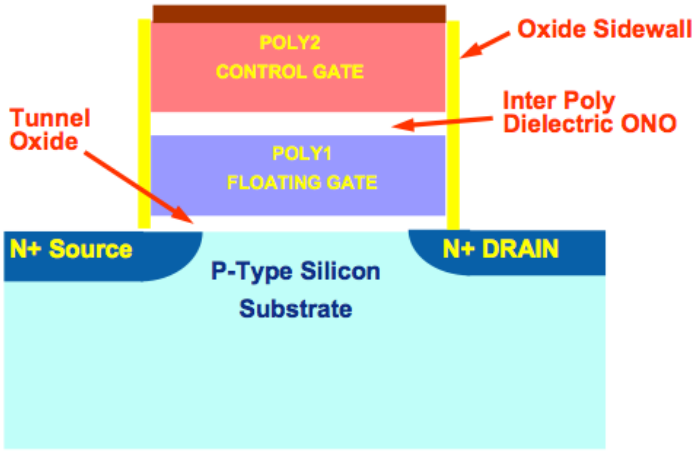


Figura 1.3: *Rappresentazione schematica della struttura di un transistor floating gate. Questo dispositivo rappresenta la cella elementare di una memoria FLASH.*

Oggigiorno, lo strato di ossido tra FG e substrato, chiamato anche *tunnel dielectric*, ha uno spessore t_1 variabile tra gli 8 e 11 nm. Riguardo allo strato di ossido tra il FG ed il contatto di *gate* accessibile dall'esterno, chiamato *control gate* (CG), è normalmente costituito da una sovrapposizione di più strati di ossido e nitruro, *Oxide–Nitride–Oxide* (ONO), ed è chiamato *control dielectric*. La qualità richiesta da questo isolamento è meno critica rispetto a quella del *tunnel dielectric*. Infatti, l'unica specifica da assicurare è quella di un buon isolamento elettrico del FG durante la fase di ritenzione del dato. Non è richiesto alcun passaggio di carica attraverso questo *dielectric layer* durante le altre operazini della cella, quindi può essere adoperato uno spessore maggiore. Tipicamente, l'*Equivalent Oxide Thickness* (EOT), *i.e.* lo spessore equivalente di SiO_2 che presenta la stessa capacità, t_2, è di circa 15 nm.

Lo scopo di una struttura di *gate* siffatta (vedi Fig. 1.3) è di immagazzinare della carica elettrica nel FG e usare questo tipo di transitior come elemento di memoria. Pertanto, sono possibili due situazioni:

Nessuna carica immagazzinata nel FG ($Q_{FG} = 0$): l'inversione nel canale è raggiunta quando il voltaggio del *control gate* è uguale alla tensione di soglia della struttura M–O–S con uno spessore totale dell'ossido uguale a t_1+t_2, in riferimento, per semplicità, ad una struttura monodimensionale come rappresentato in Fig. 1.4. In questa condizione due differenti campi elettrici F_1^0 e F_2^0 sono presenti nel dielettrico di tunnel e di controllo, rispettivamente.

Carica negativa immagazzinata nel FG ($Q_{FG} < 0$): l'inversione nel canale

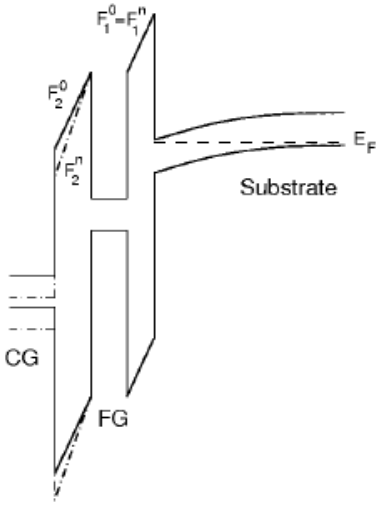


Figura 1.4: *Struttura a bande lungo una sezione verticale dal control gate, attraverso gli strati di dielettrici, fino alla zona di canale, per quanto riguarda una cella di memoria FLASH del tipo illustrata in Fig. 1.3 in condizione di soglia. Le linee continue si riferiscono al caso $Q_{FG} = 0$, mentre quelle tratteggiate al caso $Q_{FG} < 0$. F_1^0 e F_2^0 sono i campi elettrici nei dielettrici di tunnel e di controllo, rispettivamente, per $Q_{FG} = 0$, mentre F_1^n e F_2^n sono i campi elettrici per $Q_{FG} < 0$.*

avviene quando il campo elettrico nel *tunnel oxide* F_1^n à uguale a quello in situazione di nessuna carica F_1^0, mentre un campo elettrico nel *control dielectric* F_2^n più alto è necessario per contrastare la carica negativa Q_{FG} all'interno del FG. Questa evenienza è illustrata in Fig. 1.4.

Per avere una miglior comprensione degli effetti della presenza di carica negativa all'interno del FG, consideriamo i campi elettrici agenti sulla struttura nelle due situazione appena descritte.

$$F_1^0 = F_1^n$$

$$F_2^0 < F_2^n$$

In particolare, in termini di variazione di tensione di *control gate* abbiamo

$$\Delta V_2 = \Delta F_2 \cdot t_2 = (F_2^n - F_2^0) \cdot t_2 = -\frac{Q_{FG}}{C_2} \tag{1.1}$$

dove C_2 è la capacità tra *control gate* e FG, e ΔV_2 rappresenta la differenza di caduta di tensione tra la situazione in cui c'è carica immagazzinata nel FG

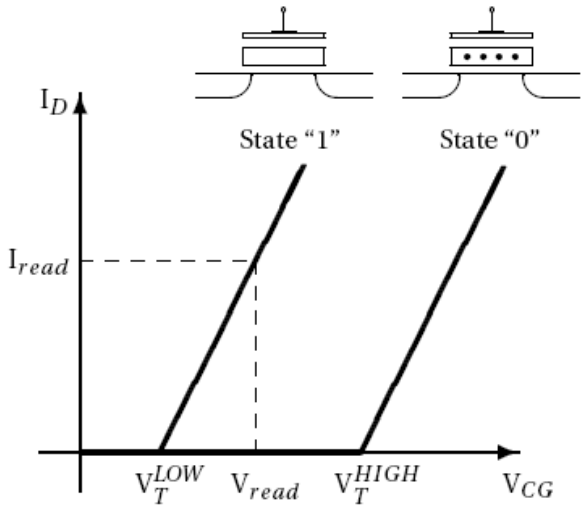


Figura 1.5: *Rappresentazione qualitativa della caratteristica $I_D - V_G$ per una cella di memoria FLASH nello stato neutro (1) e quello caricato negativamente (0).*

($Q_{FG} < 0$) e la situazione in cui non vi è carica immagazzinata ($Q_{FG} = 0$). Ora è possibile calcolare la tensione del *control gate* che deve essere applicata per raggiungere la condizione di soglia del transistor

$$V_T^n = V_T^0 - F_2^0 \cdot t_2 + F_2^n \cdot t_2 = V_T^0 + t_2 \cdot \Delta F_2 \tag{1.2}$$

Sostituendo l'Eq. 1.1 nell'Eq. 1.2 la tensione di soglia della cella risulta essere

$$V_T^n = V_T^0 - \frac{Q_{FG}}{C_2}$$

Quindi, la variazione della tensione di soglia ΔV_T può essere scritta come

$$\Delta V_T = V_T^n - V_T^0 = -\frac{Q_{FG}}{C_2} \tag{1.3}$$

Il significato dell'Eq. 1.3 sta nel fatto che la presenza di una carica negativa immagazzinata nell'elettrodo FG provoca uno *shift* positivo della tensione di soglia del transistor. Questo significa che sono necessari voltaggi di *gate* più elevati per raggiungere la condizione di soglia. Se nel FG fosse invece immagazzinata della carica positiva, l'Eq. 1.3 continuerebbe a valere, indicando come lo *shift* di tensione sarebbe questa volta negativo.

In Fig. 1.5 è rappresentata schematicamente la caratteristica corrente di *drain*–tensione di *gate*, *i.e.* $I_D(V_G)$, relazione per la struttura di Fig. 1.3 in

assenza di carica, $Q_{FG} = 0$, e in presenza, $Q_{FG} < 0$, di una certa quantità di carica negativa. La differenza tra il valore di tensione di soglia risultante dalla presenza di carica nel FG causa un diverso stato di conduzione per la cella, quando al *control gate* è applicato un certo *range* di tensione. In particolare, se è applicata una tensione di *control gate* intermedia tra lo stato a bassa tensione di soglia, V_T^{LOW}, e quello ad alta tensione di soglia, V_T^{HIGH} , per il *sensing* della corrente, è possibile osservare una netta discriminazione dello stato della cella. Riferendosi sempre alla Fig. 1.5, quando la corrente scorre attraverso il dispositivo, *i.e.* la tensione di soglia è bassa, la cella è nello stato *erased*, convenzionalmente indicato come lo stato logico "1". Quando la corrente invece non scorre, *i.e.* la tensione di soglia è alta, la cella è nello stato *programmed*, convenzionalmente indicata come lo stato logico "0".

§1.2 Nel cuore dell'elettrostatica del dispositivo floating gate

Il FG è l'elettrodo che fisicamente controlla la conduttività del canale del transistor, per questo è importante conoscere con precisione il suo potenziale elettrico. Dal momento che il FG è completamente circondato da materiali dielettrici isolanti, vi è una capacità di accoppiamento con ciascuno di essi che può influenzare il suo potenziale. Pertanto è necessario tenere in conto l'effetto dei terminali di *source*, di *drain*, di *control gate* e di *bulk* (anche chiamato substrato). Questo è un fatto di sempre maggior impatto se consideriamo la miniaturizzazione delle dimensioni della cella: le capacità parassite aumentano e con esse l'interferenza elettrica.

Per affrontare un'analisi accurata, possiamo riferirci alla sezione schematica mostrata in Fig. 1.6 dove sono enfatizzate le capacità parassite tra i terminali. Di seguito, C_{CG}, C_S, C_D e C_B sono le capacità tra il FG ed il *control gate*, *source*, *drain* e il *bulk*, rispettivamente (in questo caso C_{CG} e C_B sono le C_2 e la C_1 della sezione precedente, rispettivamente). Di conseguenza, il potenziale di FG può essere scritto come

$$V_{FG} = \frac{C_{CG}}{C_T} \cdot V_{CG} + \frac{C_S}{C_T} \cdot V_S + \frac{C_D}{C_T} \cdot V_D + \frac{C_B}{C_T} \cdot V_B + \frac{Q_{FG}}{C_T} \qquad (1.4)$$

dove V_{CG}, V_S, V_D e V_B sono i potenziali di *control gate*, *source*, *drain* e di *bulk*. Q_{FG} è la carica immagazzinata nel FG, mentre la capacità totale è $C_T = C_{CG} + C_S + C_D + C_B$.

Possiamo osservare nell'Eq. 1.4 come il V_{FG} non dipenda solo dal potenziale di *control gate*, ma anche da quello di *source*, di *drain* e di *bulk*. Consideriamo la situazione in cui il *source* ed il *bulk* siano entrambi al potenziale di riferimento, *i.e.* a massa, l'Eq. 1.4 può essere scritta come

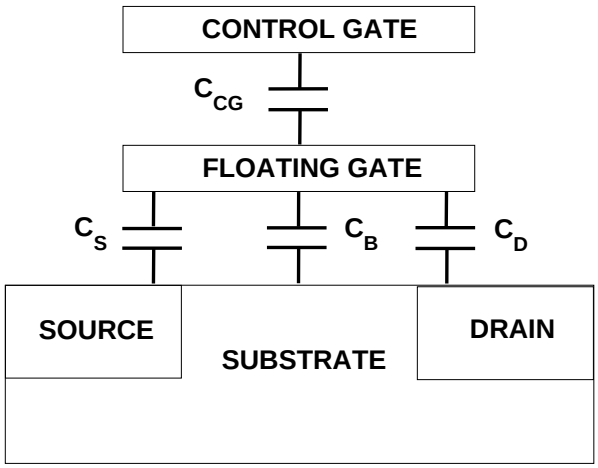


Figura 1.6: *Capacità di accoppiamento in una cella di memoria FLASH tra il FG e gli altri terminali del dispositivo*

$$V_{FG} = \frac{C_{CG}}{C_T} \cdot V_{CG} + \frac{C_D}{C_T} \cdot V_{DS} + \frac{Q_{FG}}{C_T} \tag{1.5}$$

dove possiamo isolare due rapporti che rendono conto dell'ammontare dell'influenza del *control gate* e del *drain* sul potenziale del FG, chiamati *coupling factors* e così definiti

$$\alpha_G = \frac{C_{CG}}{C_T} = \frac{C_{CG}}{C_{CG} + C_S + C_D + C_B} \tag{1.6}$$

$$\alpha_D = \frac{C_D}{C_T} = \frac{C_D}{C_{CG} + C_S + C_D + C_B} \tag{1.7}$$

Usando le definizioni dell'Eqs. 1.6 e 1.7 nell'Eq. 1.5

$$V_{FG} = \alpha_G \cdot V_{CG} + \alpha_D \cdot V_{DS} + \frac{Q_{FG}}{C_T} \tag{1.8}$$

Un'accurata analisi dei *coupling factors* α_G e α_D mostrano come questi parametri dipendano dalla polarizzazione della cella [1]. Infatti, dalle Eq. 1.6 e Eq. 1.7 è evidente la dipendenza da C_B, che rappresenta l'accoppiamento del FG con il substrato. In condizioni di lavoro al di sotto della tensione di soglia, l'accoppiamento è dovuto alla serie della capacità del *tunnel oxide* C_{tun} e della capacità del substrato in inversione di popolazione C_{dep}

$$\frac{1}{C_B} = \frac{1}{C_{tun}} + \frac{1}{C_{dep}}$$

Al di sopra della tensione di soglia, invece, C_B corrisponde solamente alla capacità del *tunnel oxide*, che è C_{tun}. In questa situazione, α_G risulta essere più bassa che non in condizione di sottosoglia. Valori tipici di α_G sono compresi tra 0.55 e 0.65 in condizione di sopra soglia, mentre possono raggiungere valori fino a 0.7 in caso di sottosoglia. Per quanto riguarda α_D, invece, i suoi valori tipicamente si aggirano attorno ai $0.1 \div 0.2$.

Come visto in precedenza, il parametro più importante di un transistore FG è la tensione di soglia. Questa è il potenziale $V_{T_{FG}}$ che bisogna applicare al FG, con $V_{DS} = 0$, per raggiungere l'inversione nel canale (*i.e.* la densità di portatori minoritari, alla superficie del substrato, eguaglia la concentrazione di droganti). Dal momento che il FG è un terminale completamente isolato dal resto del dispositivo attraverso materiali isolanti, non è accessibile in maniera diretta e quindi per raggiungere la condizione di inversione deve essere applicato un potenziale $V_{T_{CG}}$ adatto, derivato dall'Eq. 1.8

$$V_{T_{CG}} = \frac{1}{\alpha_G} \cdot V_{T_{FG}} - \frac{Q_{FG}}{C_{CG}} \tag{1.9}$$

Interpretando l'Eq. 1.8 possiamo notare come, supponendo $V_{CG} = 0$, $V_{T_{FG}}$ dipenda solo dalla teconologia del dispositivo (α_G e C_{CG}), mentre $V_{T_{CG}}$ varii anche con la carica intrappolata nel FG. Questa proprietà può essere utilizzata per immagazzinare dell'informazione in maniera non–volatile. Infatti, controllando il valore di $|Q_{FG}/C_{CG}|$, è possibile ottenere un'opportuno *shift* di soglia per definire due differenti e ben definiti stati della cella. Allo stato *erased* corrisponde $Q_{FG} = 0$, mentre allo stato programmato corrisponde $Q_{FG} < 0$. In termini di tensione di soglia applicata al *control gate*, otteniamo

$$V_{T_{CG}} = \tfrac{1}{\alpha_G} \cdot V_{T_{FG}} = V_{T_E}$$

$$V_{T_{CG}} = \tfrac{1}{\alpha_G} \cdot V_{T_{FG}} - \tfrac{Q_{FG}}{C_{CG}} = V_{T_P}$$

che corrispondono, rispettivamente, alla soglia cancellata e a quella programmata. Come visto in precedenza, applicando al CG una tensione intermedia (come mostrato in Fig. 1.5), è possibile discriminare lo stato della cella. La cella è nello stato di ON, *i.e.* la corrente fluisce attraverso il dispositivo, poiché non vi è nessuna carica nel FG, mentre la cella è OFF, *i.e.* la corrente non scorre attraverso il dispositivo, dal momento che vi è della carica immagazzinata nel FG.

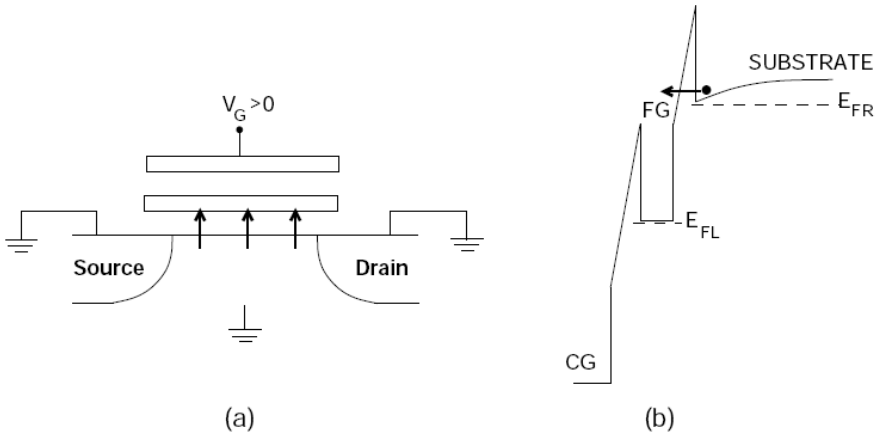


Figura 1.7: *Polarizzazione adottata per la programmazione di tipo Fowler–Nordheim (a) e corrispondente struttura a bande del dispositivo (b).*

§1.2.1 Meccanismi di programmazione/cancellazione

Come abbiamo visto, per controllare lo stato logico della cella, è necessario controllare l'immissione e l'emissione di carica elettrica verso e dal FG. L'operazione di programmazione consiste dunque nell'iniettare all'interno del FG della carica elettrica proveniente dal substrato, in modo tale alzare la tensione di soglia al CG e portare la cella nello stato logico "0", mentre l'operazione di cancellazione avviene mediante lo svuotamento della carica presente nel FG verso il substrato, riportando la tensione di soglia al CG al livello più basso, che consiste nello stato logico "1".

Esistono fondamentalmente due modi grazie ai quali posso ottenere questo flusso di elettroni attraverso il *tunnel oxide*, che sono il meccanismo di tunneling di Fowler–Nordheim e quello di *Channel Hot Electrons* (CHE). Entrambi i processi necessitano di una corretta polarizzazione della cella, in particolare del potenziale di FG, controllato dall'accoppiamento capacitivo con gli altri terminali.

Fowler–Nordheim tunneling

Il trasferimento di carica per effetto tunnel è un fenomeno prettamente quanto–meccanico. È un meccanismo molto utile in quanto il flusso di corrente richiede un basso consumo di potenza ed è tuttavia molto efficiente, dal momento che non vi è carica dispersa agli altri terminali. Sia in programmazione che in cancellazione il tunneling avviene lungo tutta la lunghezza del canale. In programmazione, il flusso netto di carica dal substrato al *gate* viene originato applicando una polarizzazione positiva al *control gate* mantenendo al potenziale

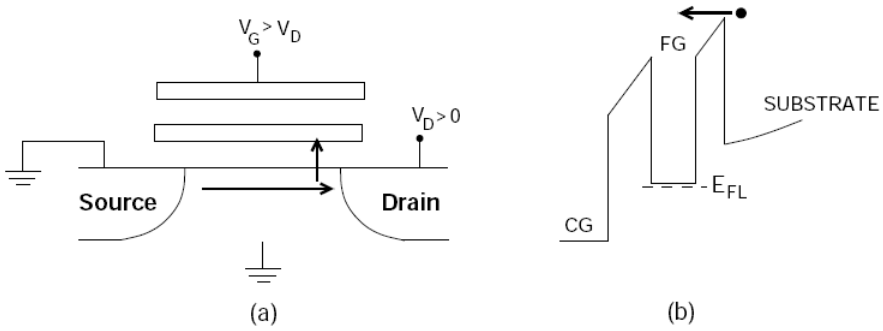


Figura 1.8: *Polarizzazione adottata per la programmazione di tipo CHE (a) e corrispondente struttura a bande del dispositivo (b).*

di riferimento gli elettrodi di *source*, *drain* e *bulk*. La cancellazione avviene in modo analogo, con l'unica differenza che ora è il *bulk* polarizzato positivamente, mentre il *control gate* viene posto a massa.

L'intensità del flusso di carica, e quindi la durata delle operazioni di programmazione/cancellazione, è determinata dalla probabilità di *tunneling* attraverso la barriera di potenziale tra substrato e *floatig gate*, probabilità che si sa dipendere esponenzialemente dal campo elettrico agente sul *tunnel oxide*. Le tensioni applicate agli elettrodi sono comunque molto elevate, allo scopo di velocizzare il più possibile le operazioni di programmazione/cancellazione. Esse pertanto si aggirano attorno ai $18V \div 20V$.
Per effetto della forte polarizzazione, la banda di conduzione della struttura risulta di pendenza tale che la barriera di potenziale risentita dagli elettroni è sottile e di forma triangolare, come mostrato in Fig. 1.7, il che corrisponde ad un regime di tunneling chiamato di *Fowler–Nordheim*.

Un altro vantaggio di questo meccanismo, se paragonato al CHE, che vedremo tra poco, è quello di essere distribuito su tutta la lunghezza di canale del dispositivo, prevenendo dannosi effetti di degradazione locale dell'ossido di tunnel. Il principale svantaggio è, invece, la bassa intensità del flusso di carica, dilatando i tempi di programmazione/cancellazione rispetto al metodo CHE.

Channel Hot Electrons (CHE)

Il trasferimento di carica tra substrato e *floating gate* può avvenire anche tramite l'iniezione di elettroni ad elevata energia, da cui l'aggettivo *Hot*, così elevata da superare in energia il fondo della banda di conduzione del *tunnel oxide*. In Fig. 1.8 è schematizzato il processo di programmazione della cella: una tensione di *drain* di circa $4 \div 5V$ genera un campo longitudinale che aumenta l'energia cinetica dei portatori nel tragitto tra *source* e *drain*, in modo tale che una volta che gli elettroni abbiano raggiunto il *drain* hanno acquisito

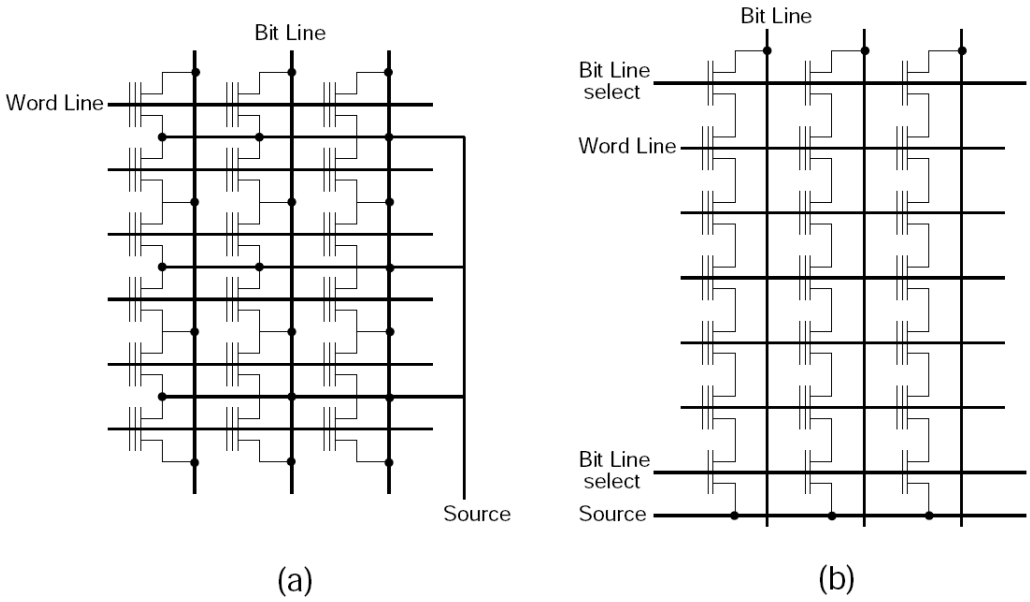


Figura 1.9: *Architetture di memorie FLASH di tipo NOR (a) e di tipo NAND (b).*

un'energia tale da poter superare la barriera mostrata dall'ossido, agevolate naturalmente dalla polarizzazione di *gate*.

Il meccanismo CHE consente di avere un flusso di corrente molto intenso, il che concede dei tempi di programmazione/cancellazione decisamente più rapidi rispetto al meccanismo Fowler–Nordheim, tuttavia, essendo il flusso concentrato in una ristretta zona spaziale – in corrispondenza del *drain* – si otterrà un maggior danneggiamento della struttura cristallina, con tutti gli svantaggi derivanti da un eccessivo stress della struttura. Un altro svantaggio è l'elevata dissipazione di potenza e le bassa efficienza di iniezione, infatti è richiesta una corrente elettrica tra *source* e *drain* di circa $10 \div 100 \mu A$.

§1.3 Architetture FLASH

Nel corso degli anni, diversi tipi di architetture sono stati sviluppati, tuttavia soltanto due di essi sono diventati capisaldi di questa tecnologia: l'architettura NOR a massa comune e l'architettura NAND (Fig. 1.9). Il nome di ciascuna architettura deriva dalla configurazione topologica dell'array. Infatti, nell'architettura NOR le celle sono connesse le une alle altre in parallelo, come i transitori della rete di *pull–down* di una porta NOR in logica CMOS. Nell'architettura NAND, invece, le celle sono connesse in serie, per il motivo

NOR FLASH	Programmazione	Cancellazione
V_{WL}	$8 - 10V$	$-8V$
V_{BL}	$4 - 5V$	$6 - 8V$
T_{Pulse}	$10\mu s$	$\sim 1s$
I_{Drain}	$10 - 100\mu A$	≈ 0

Tabella 1.1: *Polarizzazioni adottate per la programmazione e la cancellazione di un array di tipo NOR. La cancellazione è effettuata mantenendo i contatti di drain, source e substrato allo stesso potenziale.*

analogo al precedente. Descriviamo ora brevemente le caratteristiche salienti di ciascuna delle due architetture appena introdotte.

L'architettura NOR consente un rapido accesso ai dati ed impiega il metodo di programmazione CHE, mentre la cancellazione avviene tramite il metodo di tunneling FN. La struttura NAND invece utilizza il tunneling FN sia per la programmazione che per la cancellazione e presenta dei tempi di scrittura e lettura più lunghi rispetto all'altra architettura. Nella configurazione NOR le celle di memoria sono connesse da *word lines* e *bit lines* secondo lo schema in Fig. 1.9(a), e come si vede le celle condividono a coppie lo stesso contatto di *drain* e di *source*. Il valore della corrente di lettura si aggira attorno ai $6\mu A$. Nella configurazione NAND le celle invece sono connesse alle *word lines* e alle *bit lines* come mostrato in Fig. 1.9(b) e formano catene di 16 o 32 celle. Grazie all'assenza di contatti di drain all'interno della catena, le dimensioni sono molto più compatte rispetto alle NOR. Gli svantaggi principali però risiedono nella bassa corrente di lettura, attorno ai $500nA$, che allunga i tempi di accesso, e nell'altrettanto lenta corrente di scrittura. L'impossibilità di accedere direttamente al terminale di *drain* della singola cella impedisce di fatto l'utilizzo della tecnica di programmazione CHE.

NAND FLASH	Programmazione	Cancellazione
V_{WL}	$18 - 20V$	$0V$
V_{Bulk}	$0V$	$18 - 20$
T_{Pulse}	$20\mu s$	$1ms$
I_{Drain}	≈ 0	≈ 0

Tabella 1.2: *Polarizzazioni adottate per la programmazione e la cancellazione di un array di tipo NAND. La cancellazione è effettuata mantenendo i contatti di drain, source e substrato allo stesso potenziale.*

Le differenti caratteristiche delle due architetture, riassunte nelle Tabs. 1.1 e 1.2, le rendono adatte ad ambiti diversi. La topologia NOR, infatti, veloce ma

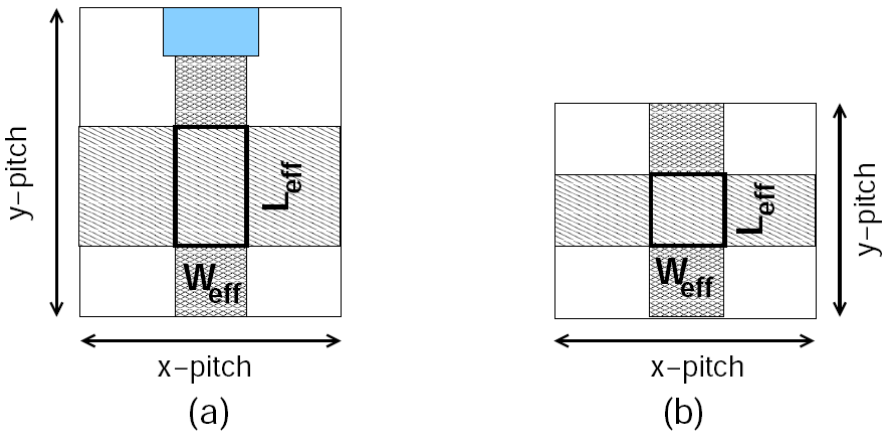


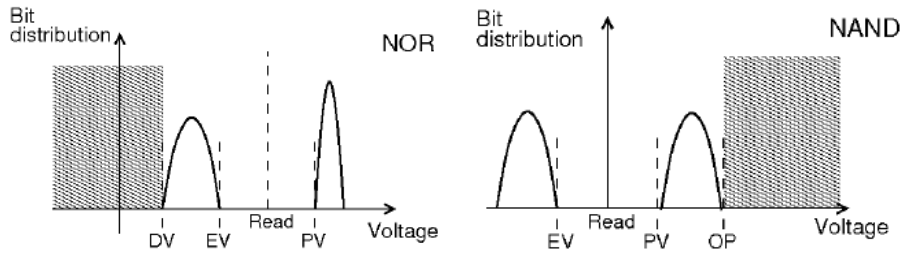
Figura 1.10: *Layout schematico di una cella FLASH di tipo NOR (a) di tipo NAND (b).*

dal maggior consumo di potenza, viene utilizzata per applicazioni *code storage*, *i.e.* memorizzazione di programmi , BIOS e simili; la topologia NAND, invece, più efficiente ma meno rapida, viene utilizzata per applicazioni *data storage*, *i.e.* memorie di massa a stato solido.

La Fig. 1.10 schematizza il *layout* di una cella NOR (a) e di una cella NAND (b). La struttura è sostanzialmente rimasta invariata, nel corso degli anni e delle varie generazioni tecnologiche sinora adottate. La principale differenza tra le due architetture sta nella dimensione indicata con *y–pitch*. Infatti, la configurazione NOR richiede un contatto di *drain* condiviso a due a due dalle celle, contatto assente nella configurazione NAND. Inoltre, la tecnologia NOR necessita di una lunghezza di canale maggiore, al fine di evitare fenomeni di *punch–through* durante la programmazione CHE. La lunghezza lungo y è pertanto maggiore per le NOR che non per le NAND, e questo spiega la maggiore dimensione della cella del primo tipo rispetto alla cella del secondo.

Nelle Figs. 1.11(a) e 1.11(b) sono mostrate le distribuzioni delle tensioni soglia per le due diverse architetture. Come si vede, il fatto di avere delle celle in serie o in parallelo impatta fortemente sulla scelta delle tensioni di soglia *erased* e *programmed*. In particolare, si evince come per le NAND la soglia *erased* sia negativa, accorgimento necessario per la lettura del dato lungo la *bit line* quando la tensione di lettura è $V_{CG} = 0V$, fenomeno che invece non accade per le NOR, che essendo lette "singolarmente" hanno tutte le tensioni di soglia maggiori di zero.

L'evoluzione della tecnologia FLASH per le memorie non volatili è tesa ad ottenere densità di integrazione sempre più elevate e al contempo di migliorare le prestazioni dell'*array*. Come per la tecnologia CMOS in generale, si perseguono questi obiettivi attraverso la riduzione delle dimensioni dei dispositivi,



(a) (b)

Figura 1.11: *Distribuzione della tensione di soglia per una cella di tipo NOR (a) e NAND (b). Le aree ombreggiate devono essere evitate, dal momento che causano errate letture. (a): EV=Erase Verify level, PV=Program Verify level, DV=Depletion Verify level; (b): EV=Erase Verify level, PV=Program Verify level, OP=Over Programming level.*

i.e. lo *scaling*. Il principale limite allo *scaling* delle celle FLASH è costituito dai problemi di affidabilità imputabili al *tunnel oxide*. Il suo spessore e la sua qualità sono, infatti, critici per quanto riguarda il mantenimento del dato memorizzato. Le cause di perdita di carica dal FG attraverso il *tunnel oxide* sono essenzialmente due: il *tunneling* quantistico, che è un fenomeno intrinseco, e lo *Stress Induced Leakage Current* (SILC), che invece è un fenomeno estrinseco. Entrambi i meccanismi aumentano di intensità al diminuire dello spessore dell'ossido, poiché la trasparenza di barriera aumenta in modo esponenziale. L'insorgere del SILC è legato alla generazione di stati trappola (difetti reticolari, impurità varie) nel *gap* energetico proibito del *tunnel oxide* durante le operazioni di programmazione/cancellazione, operazioni durante le quali il flusso di carica che viene fatto scorrere nell'ossido danneggia la struttura cristallina di quest'ultimo. Esiste quindi, per esso, uno spessore minimo tale da garantire una ritenzione del dato per un tempo sufficientemente lungo, tipicamente dieci anni. Allo stato attuale, gli spessori minimi di SiO_2 sono attorno ai $9nm$ per le celle NOR e attorno ai $7nm$ per le celle NAND. Lo spessore delle celle NOR è più alto in considerazione del metodo di programmazione: il flusso di elettroni "caldi" sottopone il dielettrico di tunnel ad un maggior degrado, aumentanto la probabilità del SILC.

La Fig. 1.12 mostra il tipico andamento della corrente di perdita causato dal SILC, in funzione del potenziale di *floating gate*. Le celle affette da SILC presentano una corrente a basse tensioni decisamente più elevate rispetto a quelle per le quali il meccanismo di perdita è sostanzialmente quello di *tunneling* intrinseco.

Si pensa che il modo più indicato per rilassare il vincolo sullo spessore dell'ossido, sia quello di utilizzare materiali dielettrici diversi dal SiO_2, a più

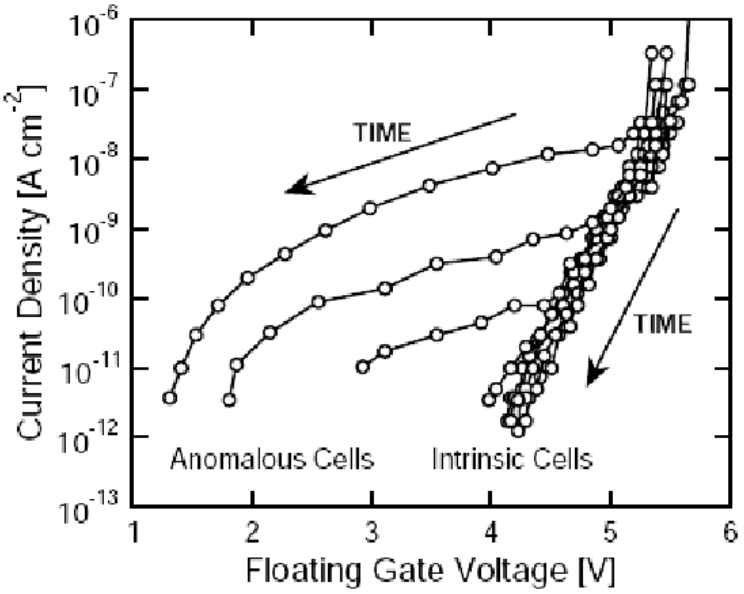


Figura 1.12: *Densità di corrente al variare della tensione di floating gate per celle di memoria FLASH aventi spessore dell'ossido di tunnel di* $6.5nm$*. Sono mostrati il comportamento anomalo e quello intrinseco. Le caratteristiche sono estratte da misure su celle cancellate e sottoposte ad uno stress positivo con una tensione di control gate di* $7.5V$ *[3].*

alta costante dielettrica ϵ_{HK}. Per tale motivo sono chiamati dielettrici *High–k* (HK) [4]. L'impossibilià di ridurre lo spessore degli isolamenti, senza cambiare materiale dielettrico, ostacola pesantemente lo *scaling* delle dimensioni planari, il cui rapporto di forma aumenta al diminuire del nodo tecnologico.

§1.4 Uno sguardo alla tecnologia High–κ

Come già accennato in precedenza, lo *scaling* aggressivo dei dispositivi CMOS ha messo in luce, nell'ultima decade, un problema di natura fondamentale per quanto riguarda l'ossido di *gate*, e cioè che per gli spessori richiesti dalla miniaturizzazione la corrente di *gate*, che è una corrente indesiderata (come si è visto per il caso delle memorie FLASH), cresce fino a valori inaccettabili. Questo perchè la corrente di *leakage* è un fenomeno legato al tunneling diretto/Fowler–Nordheim, che come si sa è un meccanismo di trasporto intrinseco. La densità di corrente per gli elettroni ha la forma funzionale del tipo

$$J = Ae^{\left(-\frac{B\Delta E_c^n d}{V}\right)} \tag{1.10}$$

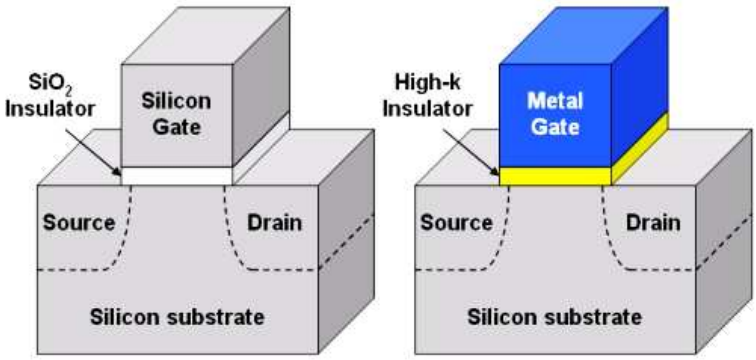


Figura 1.13: *Rappresentazione schematica della struttura di un MOSFET con dielettrico di gate di SiO_2 (a) e di un materiale High–k (b).*

dove ΔE_c^n rappresenta l'*offset* della banda di conduzione tra substrato e dielettrico, V è la caduta di tensione nel materiale, d è lo spessore e B_n è una costante. Il fattore A contiene anch'esso il campo elettrico V/d, ma possiamo, ai nostri fini, considerarlo una costante, dal momento che una dipendenza dal campo elettrico è presente anche all'esponente, dominando di fatto su A. Il fattore B invece include la massa efficace ai fini della conduzione dell'elettrone nel materiale. Per una prima stima qualitativa si può considerare anch'esso come una costante. È evidente dunque che se d diminuisce, a seguito dello *scaling* dello spessore dell'ossido, la densità di corrente aumenta in modo esponenziale, considerando che che V non scala tanto quanto le dimensioni fisiche, *i.e.* il campo elettrico nel materiale aumenta. L'idea alla base della tecnologia *High–k* è pertanto quella di sostiuire il classico SiO_2 con un dielettrico a più elevata permettività dielettrica $\kappa_{HK} > \kappa_{ox}$, da cui il nome (per essere coerenti con il resto del lavoro, tuttavia, la permettività dielettrica verrà indicata con ϵ), in maniera tale che a parità di capacità tra substrato e *gate* ci si possa concedere un $d_{HK} > d_{ox}$, e di conseguenza $J_{HK} < Jox$. Il motivo per cui si mantiene inalterata la capacità del dielettrico è che in questo modo non si va ad alterare l'elettrostatica del substrato (a parità ovviamente di tensione di *gate*), che è già stata ottimizzata per altre vie seguendo le regole dello *scaling*. In Fig. 1.13 possiamo apprezzare il passaggio da un isolante di *gate* di normale diossido di silicio ad un generico materiale *High–k*. Trovandoci quindi in presenza di un dielettrico ad alta permettività elettrica e tenendo presente l'ultima osservazione

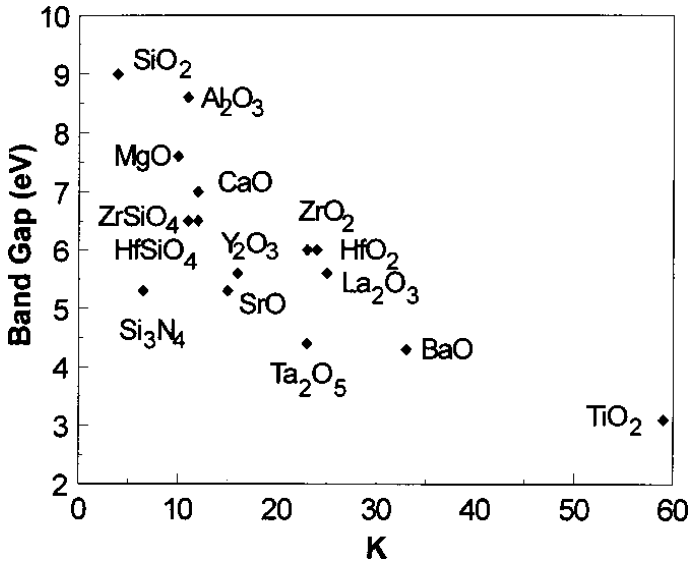


Figura 1.14: *offset delle bande di conduzione per differenti ossidi in funzione del valore* ϵ_{HK} *[8].*

$$C_{HK} = C_{ox}$$

$$\frac{\epsilon_{HK}}{t_{HK}} = \frac{\epsilon_{ox}}{t_{ox}}$$

$$t_{ox} = t_{HK}\frac{\epsilon_{ox}}{\epsilon_{HK}}$$

dove il t_{ox} corrisponde allo spessore di SiO_2 equivalente, dal punto di vista elettico, al t_{HK}, e prende il nome di *Equivalent Oxide Thickness* (EOT).

I requisiti per questi materiali di nuova concezioni sono tuttavia molto rigidi. Accanto all'assoluto bisogno di un'elevata permettività dielettrica ϵ_{HK}, bisogna tenere in considerazione che cambiando materiale cambia anche un altro parametro fondamentale nell'Eq. 1.10: l'*offset* della banda di conduzione tra substrato e dielettrico ΔE_c. Sfortunatamente è stato mostrato [5], attraverso considerazioni sulla fisica dei legami all'interno del reticolo cristallino, come l'aumentare della ϵ_{HK} implichi una diminuzione del ΔE_c. Questa diminuzione è contraria ai criteri che sottendono la tecnologia *High–k*, dal momento che tenderebbe a far aumentare la corrente di tunnel. La situazione è evidente in Fig. 1.14, dove sono indicati gli *offset* delle bande di conduzione in funzione della permettività dielettrica del materiale.
Si tratta quindi di trovare un materiale che abbia un buon compromesso tra ϵ_{HK} e ΔE_c, in modo da soddisfare comunque la richiesta di una minore corrente di *leakage*.

Un altro problema che affligge questa tecnologia è che non sempre è così

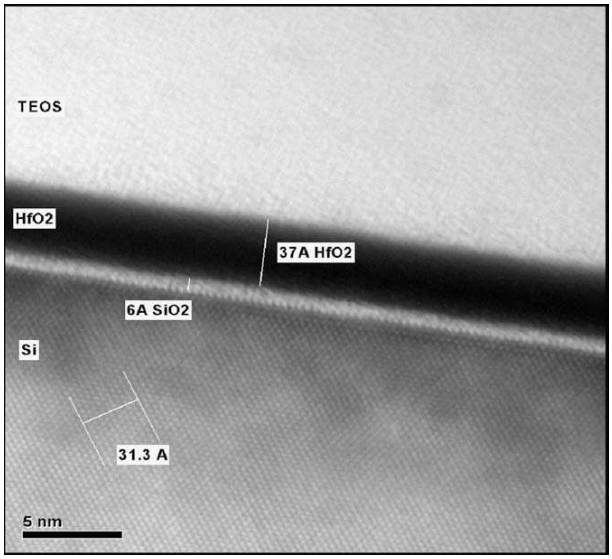


Figura 1.15: *Formazione di SiO_2 nativo all'interfaccia a seguito del processo di deposizione del dielettrico High–k [6]. È ben visibile lo strato più chiaro di diossido di silicio, dello spessore di 6Å, tra il substrato e l'ossido di afnio.*

semplice far crescere un materiale diverso dal SiO_2 sulla sperficie del substrato, che rimane sempre di silicio. La grande compatibilità tra il silicio e il suo ossido ha, nel corso della storia dell'elettronica, reso vincente questo semiconduttore sui concorrenti, e ora che al silicio non si può più rinunciare, almeno per quanto riguarda la tecnologia CMOS, ci si scontra con dei problemi di compatibilità tra materiali tutt'altro che banali. Innanzi tutto vi è un problema di *lattice matching*, ovvero della diversa dimensione dei passi reticolari tra substrato e materiale innovativo. Questo disparità geometrica si ripercuote sulla qualità dell'interfaccia $Si - HK$, che presenta infatti una bassa stabilità termodinamica ed elettrica, un'alta densità superficiale di difetti che provoca una forte diminuzione della mobilità dei portatori lungo il canale, nonché una possibile più alta densità di difetti reticolari all'interno del dielettrico stesso, come conseguenza dello *stress* fisico cui sono sottoposti i legami atomici [6].
La natura del processo chimico/fisico di deposizione, tuttavia, è tale per cui nella maggior parte dei casi viene comunque a formarsi un sottile strato di SiO_2 nativo tra substrato e HK, spesso non più di un paio di nanometri. Se da un lato questo favorisce la transizione reticolare tra i vari materiali, ripristinando in parte la qualità dell'interfaccia (tanto che a volte lo si lascia crescere di proposito), dall'altro diminuisce il vantaggio di utilizzare un materiale HK come dielettrico di *gate*, dal momento che dal suo spessore di ossido equivalente bisogna decurtare quello di ossido nativo

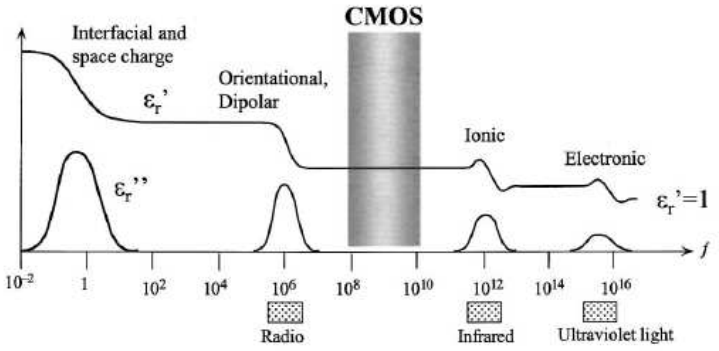


Figura 1.16: *Dipendenza dalla frequenza della parte reale (ϵ_r') e immaginaria (ϵ_r'') della permettività dielettrica. Nel range di frequenze tipico della tecnologia CMOS, essa è dominata dal contributo ionico ed elettronico [6].*

$$EOT = t_{native\ oxide} + \frac{\epsilon_{ox}}{\epsilon_{HK}} t_{HK}$$

neutralizzando parzialmente i vantaggi dell'utilizzo di un ossido HK per quanto concerne la corrente *leakage*. In Fig. 1.15 è mostrata una fotografia della sezione verticale di uno *stack* di *gate*, dove è possibile individuare, tra il substrato e l'ossido HK, lo strato di ossido di silicio nativo. I diversi spessori dipendono dal processo chimico/fisico utilizzato per la deposizione.

Numerosi altri problemi vengono alla luce, approfondendo le tematiche relative a questa tecnologia. Essi riguardano sia la caratterizzazione dei parametri fisici relativi materiali HK (in Fig. 1.16 è mostrato per esempio l'andamento delle permettività dielettriche, nelle loro componenti reale e immaginaria, al variare della frequenza del campo elettrico applicato, con l'indicazione del fenomeno fisico da cui dipendono), sia lo studio della compatibiltà di questi materiali con il substrato (come accennato in precedenza), con il *gate*, la loro stabilità termodinamica ed elettrica, la morfologia del film deposto, la dipendenza di tutti questi parametri dal processo di crescita [6]... Tutti argomenti che esulano dallo scopo di questa sezione, rimandando il lettore interessato ai riferimenti bibliografici per un maggior approfondimento.

In ultimo, forniamo una tabella con le proprietà salienti dei più promettenti candidati *High–k* messi a paragone con il SiO_2 (Tab. 1.3)[6][7].

Materiale	ϵ_r	***Band Gap*** $[eV]$	ΔEc $[eV]$ **verso** ***Si***	**Struttura Cristallina**
SiO_2	3.9	8.9	3.2	Amorfo
Si_3N_4	7	5.1	2	Amorfo
Al_2O_3	9	8.7	2.8	Amorfo
Y_2O_3	15	5.6	2.3	Cubico
La_2O_3	30	4.3	2.3	Esagonale, Cubico
Ta_2O_5	26	4.5	1-1.15	Ortorombico
TiO_2	80	3.5	1.2	Tetragonale
HfO_2	25	5.7	1.5	Monoclino, Tetragonale, Cubico
ZrO_2	25	7.8	1.4	Monoclino, Tetragonale, Cubico
Gd_2O_3	13.6	6.4	3.1	Cubico

Tabella 1.3: *Proprietà fisiche ed elettriche dei principali candidati materiali High–k.*

§1.5 Obiettivi del lavoro di tesi

La breve rassegna sulle memorie FLASH e sulla pratica dello *scaling* descritta in questo capitolo evidenzia come la corrente di perdita di *gate* sia uno degli ostacoli principali all'ulteriore sviluppo della tecnologia CMOS. La corrente di *leakage* comporta una dissipazione di potenza supplementare nei circuiti logici e causa grossi problemi di affidabilità per le celle di memoria FLASH. Questi aspetti complementari impongono un limite inferiore molto stringente allo spessore dei dielettrici di isolamento. È quindi fondamentale modellizzare in modo accurato i meccanismi di conduzione attraverso film sottili di isolanti, al fine di comprenderne la natura e possibilmente individuare nuove soluzioni.
Nel presente lavoro si sono, pertanto, studiati ed implementati numericamente dei modelli di conduzione per *tunneling* quantistico assistito da difetto, fenomeno che consente di spiegare le correnti anomale SILC osservate nelle misure sperimentali. Inoltre, si sono implementate delle tecniche Monte Carlo che, sfruttando i precedenti modelli, fossero in grado di valutare le statitiche di diverse grandezze di interesse, al fine di estrapolare una descrizione probabilistica delle distribuzioni delle trappole nei dielettrici e una valutazione delle caratteristiche di ritenzione di celle FLASH, elementi fondamentali su cui basare i parametri di merito per una determinata tecnologia.

Capitolo **2**

Modelli di conduzione di corrente attraverso dielettrici di gate

Die Quantenmechanik gefällt mir nicht
und es tut mir leid damit zu tun zu haben.

(La meccanica quantistica non mi piace,
e mi spiace di averci avuto a che fare.)

Erwin Schrödinger

Nel presente capitolo si descriverà in modo dettagliato il processo di tunneling *quantistico ed i metodi di calcolo delle grandezze ad esso associate: le trasparenze di barriera e i tassi di* tunneling. *Si passerà poi a valutare numericamente le correnti di perdita per* tunneling *diretto/Fowler–Nordheim, 1TAT e 2TAT attraverso strutture MOS e* High–k *monodimensionali, confrontando infine i dati con alcune misure sperimentali.*

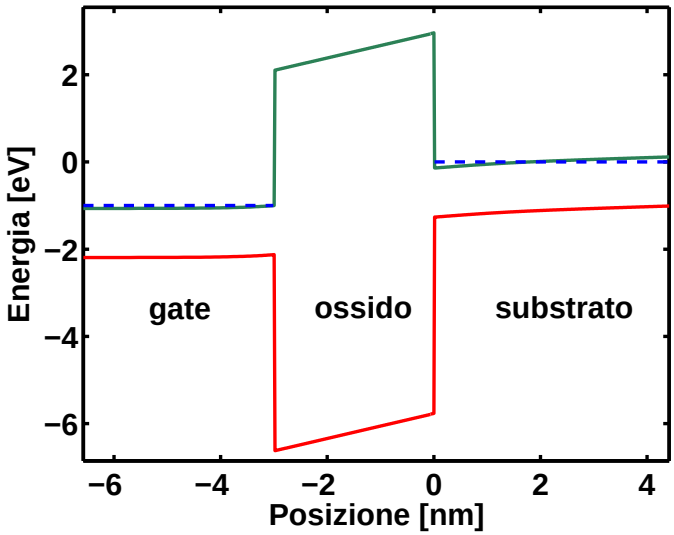


Figura 2.1: *Andamento spaziale del fondo della banda di conduzione (verde) e della cima della banda di valenza (rosso) di una struttura MOS polarizzata con $V_g = 1V$. Sono anche indicati i livelli di Fermi nel substrato e nel gate (blu).*

§2.1 Corrente di tunneling diretto e Fowler–Nordheim

§2.1.1 Introduzione al concetto di tunneling

Si consideri, innanzi tutto, una struttura MOS con substrato di tipo p e *gate* in polisilicio di tipo n^+. Applicando una tensione positiva all'elettrodo di *gate* rispetto al *bulk* ha luogo uno schema a bande come rappresentato in Fig. 2.1. Un elettrone nella banda di conduzione del substrato ha una probabilità non–nulla di attraversare, per effetto tunnel, lo strato di ossido, ed andare ad occupare uno stato libero nella banda di conduzione del *gate*. Questo corrisponde ad una corrente di "perdita" che, nelle tecnologie ultra–scalate, risulta essere di notevole entità e impatta negativamente in misura non trascurabile sui consumi di potenza e sui parametri di affidabilità.

È possibile sin da ora mettere in luce alcuni aspetti fondamentali dei processi di *tunneling*: innanzi tutto, la probabilità che un portatore attraversi la barriera di potenziale è tanto più piccola quanto più questa è spessa. Per questo le correnti di perdita sono significative solo per spessori di dielettrico di pochi nanometri, come lo è per tecnologie MOS scalate. Inoltre, più la barriera energetica è elevata rispetto all'energia del portatore, minore sarà la probabilità di attraversarla. Si assume che, durante l'attraversamento della barriera,

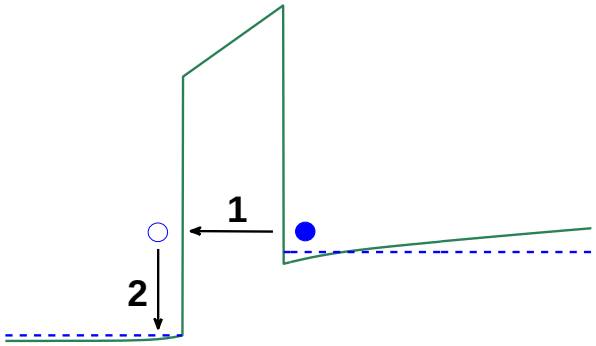


Figura 2.2: *Rappresentazione schematica della transizione per tunnel diretto di un elettrone dalla banda di conduzione del substrato a quella del gate. Sono evidenziati i due step: Il tunneling elastico attraverso la barriera e termalizzazione verso livelli energetici inferiori.*

il portatore mantenga la sua energia totale, *i.e.* la transizione è elastica, ed eventualmente rilassi, molto velocemente, solo alla fine del tragitto.

All'equilibrio termodinamico, i flussi di carica tra substrato e *gate* e viceversa si compensano, rendendo nulla corrente netta. Lo sbilanciamento dei due flussi si ottiene polarizzando la struttura MOS, in modo tale che uno acquisisca maggior vigore, mentre l'altro ne perda. Prendiamo l'esempio di Fig. 2.2 in cui è stata applicata una tensione positiva al *gate*: gli elettroni del substrato sono favoriti nel loro viaggio verso i molti stati liberi ad alta energia della banda di conduzione del polisilicio di *gate*, dove poi rilasseranno su stati a più bassa energia, *i.e.* sul fondo della banda di conduzione, dai quali un eventuale ritorno al substrato è fortemente sfavorito (l'altezza energetica della barriera è aumentata). Un ragionamento analogo si può applicare alle lacune, che in questo caso si muoveranno, in maniera netta, dalla cima della banda di valenza del polisilicio di *gate* verso gli stati vuoti a bassa energia nel substrato, per poi essere riempite da elettroni di valenza, come dire che sono naturalmente promosse a stati energetici più elevati, *i.e.* sulla cima della banda di valenza. In generale, tuttavia, il flusso di lacune è trascurabile rispetto a quello di elettroni, e ciò è ancor più vero se il dielettrico è SiO_2. Il diossido di silicio, infatti, mostra agli elettroni una barriera di circa $3.1eV$, mentre alle lacune circa $4.8eV$. Anche nel caso di utilizzo di dielettrici *High–k* il contributo degli elettroni è in genere quello dominante, dal momento che, come abbiamo visto

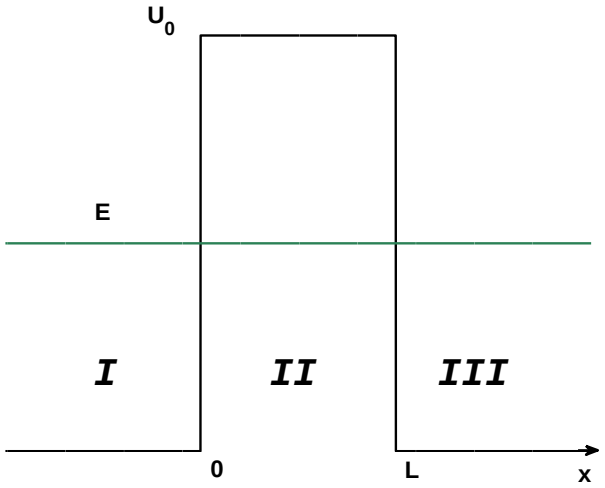


Figura 2.3: *Tunneling attraverso una barriera di potenziale rettangolare. Da notare che classicamente il moto nella regione II è proibito, per oggetti aventi energia* $E < U_0$.

nel precedente capitolo, si forma naturalmente uno strato di SiO_2 nativo tra substrato di silicio e materiale HK, riproponendo il ragionamento precedente.

Per valutare la corrente di *leakage* è necessario dare una corretta stima della probabilità di *tunneling*, detta anche *trasparenza di barriera*. Nota la trasparenza di barriera, è possibile valutare il tasso con cui i portatori attraversano il dielettrico, chiamato *tasso di transizione*. Dai tassi di *tunneling* si passa infine al calcolo del flusso di portatori che attraversa la barriera. In quest'ottica è importante valutare la corretta disposizione energetica degli stati elettronici dell'elettrodo di *gate* e di *bulk*. In particolare, vanno tenuti in conto gli effetti di quantizzazione nel canale che vanno a modificare la distribuzione energetica dei portatori stessi.

§2.1.2 Probabilità di tunneling

In accordo con il formalismo della *Envelop Function* [9], nelle strutture MOS in esame il potenziale di barriera $U(x)$ corrisponde al fondo della banda di conduzione, per gli elettroni, alla cima della banda di valenza, per le lacune. D'ora in avanti ci riferiamo solo al caso degli elettroni, per semplicità di trattazione. Una analoga trattazione tuttavia può essere utilizzata per le lacune.

Il problema ora è determinare la probabilità che un elettrone, dall'energia $E < U(x)$, attraversi la barriera di potenziale per effetto tunnel. Impostiamo il calcolo analizzando il semplice caso di una barriera dall'energia costante,

come mostrato in Fig.2.3. Questa situazione è ben lontana dal modellizzare correttamente lo strato di dielettrico di un dispositivo MOS, ma fornisce un buon punto di partenza per un'analisi più complessa, nonchè alcune indicazioni sulle dipendenze funzionali più importanti.

Il problema unidimensionale consiste in un elettrone nella regione I in moto verso la barriera. L'onda piana progressiva che descrive l'elettrone, una volta raggiunta la barriera, si divide in un'onda riflessa, che rimane nella regione I, e ad un'onda trasmessa, attraverso la regione II, alla regione III. La funzione d'onda soluzione dell'equazione di Schrödinger tempo indipendente è, in generale

$$\begin{cases} \phi_I = A_1 e^{ikx} + B_1 e^{-ikx} \\ \phi_{II} = A_2 e^{gx} + B_2 e^{-gx} \\ \phi_{III} = A_3 e^{ikx} \end{cases} \tag{2.1}$$

dove

$$k = \frac{\sqrt{2\mu E}}{\hbar}, \qquad g = \frac{\sqrt{2\mu(U_0 - E)}}{\hbar}$$

Nell'espressione si è indicata con μ la massa efficace dell'elettrone, assunta costante nelle tre regioni. A ciascuna componente euleriana del tipo Ae^{ikx} è associato un flusso di probabilità

$$F = \frac{\hbar k}{\mu}|A|^2$$

La probabilità che una particella incidente sulla barriera entri nella regione III è data dal rapporto tra il flusso incidente $F_i = \frac{\hbar k}{\mu}|A_1|^2$ e il flusso trasmesso $F_t = \frac{\hbar k}{\mu}|A_3|^2$. Analogamente, la probabilità che la particella sia riflessa è il rapporto tra il flusso riflesso $F_r = \frac{\hbar k}{\mu}|B_3|^2$. Pertanto

$$T = \frac{|A_3|^2}{|A_1|^2}, \qquad R = \frac{|B_1|^2}{|A_1|^2} = 1 - T \tag{2.2}$$

I rapporti espressi nelle Eqs. 2.2 sono determinabili imponendo la continuità delle funzioni d'onda nelle Eqs. 2.1 e delle loro derivate prime alle interfacce $x = 0$ e $x = L$. Quel che si ottiene, sviluppando i conti per la trasparenza di barriera, è

$$T = \frac{4k^2g^2}{(k^2 + g^2)\sinh^2(gL) + 4k^2g^2} \tag{2.3}$$

L'espressione può essere semplificata per $gL \gg 1$, in tal caso al denominatore $\sinh^2(gL) \approx \frac{1}{4}\exp(2gL)$, da cui

$$T \approx \frac{16k^2g^2}{(k^2+g^2)^2}\exp(-2gL) \tag{2.4}$$

La condizione $gL \gg 1$, tradotta in termini energetici, significa $E \ll U_0 - \frac{\hbar^2}{2\mu}\frac{1}{L^2}$, pertanto l'Eq. 2.4 è una valida approssimazione dell'Eq. 2.3 quando l'energia della particella è sufficientemente distante dalla sommità della barriera di potenziale.

Dall'Eq. 2.4 risulta chiaro come le dipendenze principali della probabilità di *tunneling* siano contenute nel fattore esponenziale. Il termine pre–esponenziale, invece, ha una dipendenza molto più debole dall'energia E. La trasparenza di barriera è tanto più piccola quanto più questa è alta e larga e quanto più è grande la massa della particella.

Nel caso in cui la particella incidente possieda un'energia $E > U_0$, la ϕ_{II} assume anch'essa una forma della stessa natura delle altre

$$\phi_{II} = A_2 e^{iqx} + B_2 e^{-iqx}$$

con

$$q = \frac{\sqrt{2\mu(E-U_0)}}{\hbar}$$

con il risultato che il coefficiente di trasmissione cambia leggermente forma, sostituendo $g \to iq$, ottenendo

$$T = \frac{4k^2q^2}{(k^2-q^2)\sin^2(qL) + 4k^2q^2} \tag{2.5}$$

Da notare come la trasparenza di barriera, per energie della particella superiori all'altezza della stessa, non sia sempre unitaria, come ci si aspetterebbe classicamente. In particolare, la particella ha probabilità 1 di superare la barriera solo per $q_n = n\frac{\pi}{L}$, con $n \in \mathbb{Z}^+$. Per tali valori di q, indipendenti dall'altezza di barriera e dalla massa della particella, corrisponde infatti la condizione $L = n\frac{\lambda}{2}$, che non è altro che la condizione di interferenza costruttiva delle onde in propagazione nella regione II: ai vettori d'onda risonanti $q_n = n\frac{\pi}{L}$ le onde si sommano in fase e l'euleriana trasmessa ϕ_{III} replica in ampiezza l'onda incidente ϕ_I.

Il problema che ora si pone è quello di trovare un'espressione per la trasmissività che valga, più in generale, per delle barriere di potenziale di forma arbitraria, estendendo l'interesse a risvolti più pratici. Nel corso del tempo

sono stati sviluppati diversi modi per ottenere tale risultato, ma l'approccio più noto e anche il più semplice è quello derivante dall'approssimazione semi–classica di Wentzel, Kramers e Brillouin (WKB). Altri metodi, come per esempio il *Transfer Matrix* (TM), sono più precisi e colgono delle proprietà delle funziona d'onda che alla WKB sfuggono, al prezzo di una maggiore complessità, teorica e pratica. Nel corso della presente trattazione svilupperemo nel dettaglio solo la teoria WKB, che poi sarà quella utilizzata nelle simulazioni che più avanti, in questo capitolo e nei successivi, verranno presentate.

§2.1.3 Tunneling in approssimazione semi–classica

Guidati dall'analogia formale tra l'equazione di Schrödinger e le equazioni del campo elettromagnetico, Wentzel, Kramers e Brillouin individuarono un'espressione approssimata molto raffinata per le autofunzioni dell'operatore hamiltoniano. Quest'approssimazione, in loro onore denominata WKB, è una sorta di limite in ottica geometrica dell'equazione d'onda di Schrödinger: la fase degli autostati approssimati WKB soddisfa il principio di minima azione così come, in ottica geometrica, la fase dell'onda elettromagnetica soddisfa il principio di Fermat. Per questo motivo l'approssimazione WKB viene anche chiamata "semi–classica".

Secondo il metodo WKB, la parte spaziale della generica autofunzione dell'operatore hamiltoniano assume la forma

$$\phi(x) = \frac{C_0}{\sqrt{k}} \exp\left(i \int_0^x k(x')\mathrm{d}x'\right) + \frac{C_1}{\sqrt{k}} \exp\left(-i \int_0^x k(x')\mathrm{d}x'\right) \tag{2.6}$$

dove

$$k(x) = \frac{\sqrt{2\mu(E - U(x))}}{\hbar}$$

che deriva dalla formulazione semi–classica del momento $p = \hbar k$. Nelle espressioni sono indicate con μ la massa della particella e con E l'energia dello stato.

L'espressione nell'Eq. 2.6, data dalla somma di un'euleriana progressiva e una regressiva, è molto maneggevole e può essere utilizzata efficacemente per individuare una formula generale per la trasparenza di barriera. Si può mostrare che essa costituisce un'ottima approssimazione dell'esatta funzione d'onda quando la lunghezza d'onda locale $\lambda = \frac{2\pi}{|k|}$ soddisfa la seguente

$$\left|\frac{\mathrm{d}\lambda}{\mathrm{d}x}\right| \frac{\lambda}{2\pi} \ll \lambda \tag{2.7}$$

Dal punto di vista ondulatorio, la condizione di semi–classicità si traduce nella richiesta che la variazione della lunghezza d'onda λ, nell'intervallo $\frac{\lambda}{2\pi}$, sia molto minore della lunghezza d'onda stessa. Se, per esempio, a è la dimensione caratteristica del sistema, tale che sia lecito supporre $\frac{\mathrm{d}\lambda}{\mathrm{d}x} \approx \frac{\lambda}{a}$, allora l'Eq. 2.7 diventa semplicemente $\frac{\lambda}{2\pi} \ll a$.

Tradotta in termini di meccanici di momento $p(x)$ e potenziale $U(x)$, l'Eq. 2.7 assume la forma:

$$|p^3| \gg \mu\hbar \left| \frac{\mathrm{d}U}{\mathrm{d}x} \right| \tag{2.8}$$

Da queste condizioni appare dunque evidente che l'approssimazione WKB è ragionevolmente valida in regioni spaziali in cui si hanno forti impulsi e piccoli gradienti di potenziale. In particolare, poi, essa cessa di valere intorno ai punti di classica inversione del moto, nei quali si ha $U(x) = E$, ovvero $p = 0$. Per una trattazione esaustiva si faccia riferimento a [10].

Stando a quanto detto, si può impiegare l'Eq. 2.6 per calcolare la probabilità di *tunneling* attraverso barriere di potenziale lentamente variabili nello spazio. Considerando il caso $E < U(x)$ e procedendo in maniera analoga alla precedente

$$\begin{cases} \phi_I = A_1 e^{ikx} + B_1 e^{-ikx} \\ \\ \phi_{II} = \frac{A_2}{g(x)} \exp\left(\int_0^x g(x')\mathrm{d}x'\right) + \frac{B_2}{g(x)} \exp\left(-\int_0^x g(x')\mathrm{d}x'\right) \\ \\ \phi_{III} = A_3 e^{ikx} \end{cases} \tag{2.9}$$

Dove $k = \dfrac{\sqrt{2\mu E}}{\hbar}$ e $g(x) = \dfrac{\sqrt{2\mu|U(x) - E|}}{\hbar}$. Come in precedenza, imponendo la continuità della funzione d'onda soluzione e della sua derivata prima, si possono determinare i coefficienti di trasmissione e riflessione. Per ottenere un'espressione finale della trasparenza di barriera abbastanza semplice, conviene operare alcune approssimazioni, ovviamente in coerenza con la semi–classicità.

- La derivata di $\phi_{II}(x)$ va calcolata trascurando la dipendenza da x dei termini pre–esponenziali $\frac{1}{g(x)}$. Infatti, se $U(x)$ è lentamente variabile, i termini in esame sono pressoché costanti.

- Nel sistema originato dall'applicazione delle condizioni di raccordo in $x = L$, vanno trascurati i termini esponenziali descrescenti in $\gamma = \int_0^L g(x')\mathrm{d}x'$. In accordo con l'Eq. 2.7, infatti, si ha $\gamma = \bar{g}L \gg 1$.

Dopo svariati conti si ottiene

$$T = \frac{16}{\frac{g(L)}{g(0)} + \frac{g(0)g(L)}{k^2} + \frac{k^2}{g(0)g(L)} + \frac{g(0)}{g(L)}} \exp\left(-2\int_0^L g(x')\mathrm{d}x'\right) \tag{2.10}$$

che è formalmente simile all'Eq. 2.4 e ne può essere considerata una generalizzazione, dal momento che l'Eq. 2.10 si trasforma proprio nell'Eq. 2.4 nel caso in cui $U(x) = U_0$ costante. Se la barriera di potenziale è sufficientemente alta e piatta, il fattore pre–esponenziale è dell'ordine dell'unità, consentendoci un'ulteriore approssimazione

$$T \approx \exp\left(-2\int_0^L \frac{\sqrt{2\mu(U(x') - E)}}{\hbar}\mathrm{d}x'\right) \tag{2.11}$$

L'Eq. 2.11 è quella comunemente impiegata per stimare la probabilità di *tunneling* in appossimazione WKB. A rigore, essa è valida solamente per $E < U(x)$ in tutta la regione II, ma può essere impiegata anche nel caso in cui si abbia $E < U(x)$ anche solo in un tratto della II. In quest'ultimo caso, per ottenere una buona stima della trasparenza di barriera è sufficiente svolgere l'integrale unicamente nei tratti in cui l'energia della particella è inferiore all'altezza di barriera.

Operando in questo modo, tuttavia, si perde completamente l'informazione sull'andamento della funzione d'onda per $E > U(x)$, con gli eventuali fenomeni di interferenza ad esso associati. Questo è evidentemente il limite principale dell'approssimazione WKB, ma è un prezzo che si è disposti a pagare in quanto si è comunque in possesso di un strumento molto semplice per calcolare la probabilità di *tunneling*, e la sua efficacia è comprovata dall'errore non poi così grosso nei confronti delle teorie più complesse, a meno ovviamente di non andare a considerare dei casi limite.

Premesso ciò, si può immediatamente passare ad analizzare un esempio di grande interesse pratico. In Fig. 2.4 è schematizzata una barriera di potenziale di forma trapezoidale, proprio come può essere modellizzata la barriera mostrata da uno spessore di isolante di *gate* agli elettroni (o lacune) quando ai suoi capi è applicata una differenza di potenziale. Il tratto obliquo ha pendenza $F = \dfrac{U_0 - U_1}{L}$ costante, perciò, nella regione di barriera, il potenziale assume la forma lineare $U(x) = U_0 - Fx$. Utilizzanzo allora l'Eq. 2.11 è possibile fornire un'espressione analitica per la trasparenza di barriera

$$T \approx \begin{cases} \exp\left[-\frac{4}{3}\frac{\sqrt{2\mu}}{\hbar}\frac{(U_0-E)^{3/2}-(U_1-E)^{3/2}}{F}\right] & \text{per } 0 < E < U_1, \\ \exp\left[-\frac{4}{3}\frac{\sqrt{2\mu}}{\hbar}\frac{(U_0-E)^{3/2}}{F}\right] & \text{per } U_1 < E < U_0. \end{cases} \tag{2.12}$$

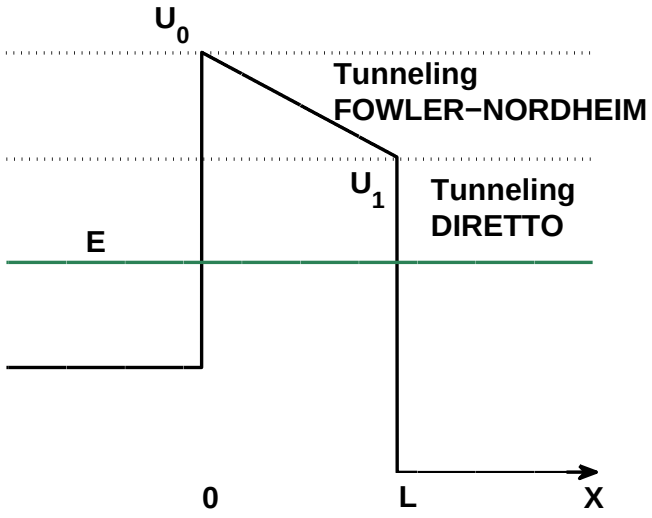


Figura 2.4: *Esempio di barriera di potenziale trapezoidale. Sono indicati i diversi regimi di tunneling a seconda dell'energia della particella incidente*

L'andamento della probabilità di *tunneling* è leggermente diverso a seconda che che l'energia della particella sia maggiore o minore di U_1. Per la prima delle Eqs. 2.12 si parla quindi di *tunnelling* di tipo *diretto*, per la seconda invece di tipo *Fowler–Nordheim*. In accordo con le approssimazioni adottate, la $T(E)$ nella regione Fowler–Nordheim non dipende in alcun modo dal tratto di barriera nel quale l'energia della particella è maggiore della barriera stessa, e per questo cresce più velocemente all'aumentare dell'energia, come risulta evidente in Fig. 2.5. L'espressione trovata mostra inoltre in modo esplicito una dipendeza esponenziale della T dal campo elettrico F costante lungo l'intera lunghezza della barriera.

L'Eq. 2.12, come abbiamo detto, è soggetta ad alcune approssimazioni che ne limitano la validità: innanzi tutto, l'aver trascurato il termine pre–esponenziale dell'Eq. 2.10 fa si che la stima fornita nella regione di tunnel diretto sia scadente per energie prossime allo 0 e a U_1. Inoltre, si è implicitamente assunto che la pendenza F sia tale da rispettare le ipotesi di approssimazione semi–classica. Per quantificare il limite di validità della teoria WKB si può utilizzare la condizione dell'Eq. 2.8, tenendo presente che $p = \sqrt{2\mu(U-E)}$. Se come valore di U scegliamo U_1, la condizione diventa

$$F \ll 2^{3/2}\frac{\sqrt{\mu}}{\hbar}(U_1 - E)^{3/2} \tag{2.13}$$

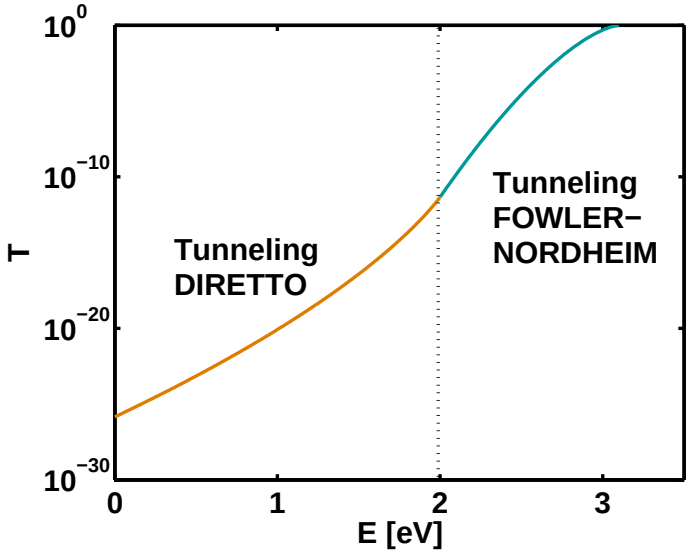


Figura 2.5: *Esempio di calcolo in approssimazione WKB della probabilità di tunneling attraverso una barriera di potenziale trapezoidale in funzione dell'energia della particella. Sono evidenziati i diversi regimi di tunneling.*

Nello studio delle correnti di perdita dei dispotivi MOS, le particelle di *tunneling* sono gli elettroni e le lacune nelle bande di conduzione e valenza, rispettivamente, per i quali $\mu \approx 10^{-31} kg$. Le barriere in questione sono alte qualche eV mentre i portatori hanno un'energia di poco superiore alla base della barriera stessa. Con tali valori, l'Eq. 2.13 fornisce il limite $F \ll 10^8 \frac{V}{cm}$, ben superiore ai normali campi elettrici che insistono sugli ossidi di gate, dell'ordine invece di $10^6 \frac{V}{cm}$. Ci aspettiamo, pertanto, che nei casi di nostro interesse la WKB coscituisca una buona approssimazione della probabilità di *tunneling*.

§2.1.4 Tassi di tunneling in approssimazione semi–classica

Supponiamo di avere una regione di controllo Ω in cui siano confinate delle particelle, e ammettiamo che esse possano sfuggire al confinamento attraversando la superficie di bordo S per effetto tunnel. Per valutare la corrente di uscita, bisogna conoscere il tasso con cui le particelle attraversano S. Nota la trasparenza di barriera T, il tasso di *tunneling* di un singolo portatore è

$$\frac{1}{\tau_{tun}} = F_{out} \cdot S = F_{in} \cdot T \cdot S \tag{2.14}$$

dove si è indicato con F_{in} il flusso di probabilità che incide sulla barriera per

unitá di area e con F_{out} il flusso in uscita da Ω. F_{out} corrisponde alla frazione di F_{in} che è stato trasmesso al di là della barriera. Assumiamo per semplicità che i flussi siano ortogonali e costanti su tutta la superficie S, e che la direzione ortogonale sia indicata con x. In generale un flusso di probabilità è definito come segue

$$F = \frac{1}{2\mu} \left(\bar{\psi}\hat{p_x}\psi - \psi\hat{p_x}\bar{\psi}\right)\Big|_{x_0} \tag{2.15}$$

dove x_0 indica la coordinata della barriera S.
Supponiamo inoltre che le dimensioni del volume di controllo siano tali da non dar luogo ad una separazione significativa dei livelli energetici, *i.e.* le autofunzioni possono essere concepite come onde piane

$$\psi(\mathbf{r}) = \frac{1}{\sqrt{SL}} e^{i\mathbf{k}\cdot\mathbf{r}} \tag{2.16}$$

dove con $\mathbf{r} = (r_x, r_y, r_z)$ si è indicato il vettore posizione e con $\mathbf{k} = (k_x, k_y, k_z)$ il vettore d'onda tridimensionale. Sostiuendo la Eq. 2.16 nell'Eq. 2.15 otteniamo

$$F_{in} = \frac{1}{SL}\frac{\hbar k_x}{\mu} \tag{2.17}$$

Ne consegue che il tasso di *tunneling* di uscita da una regione potenziale quasi–costante è

$$\frac{1}{\tau_{tun}} = T(E_x)\frac{\hbar k_x}{\mu}\frac{1}{L} \tag{2.18}$$

che come abbiamo sottolineato vale nel caso di livelli continui. Data la simmetria del sistema, inoltre, possiamo affermare che la trasparenza di barriera dipende solo dalla componente di energia nella direzione ortogonale ad S. In assenza di quantizzazione si sa che vale la relazione di dispersione parabolica

$$E_x = \frac{\hbar^2 k_x^2}{2\mu}$$

In definitiva, il tasso di attraversamento della barriera di potenziale nel caso di livelli continui dipende unicamente dall'energia nella componente ortogonale alla barriera. La probabilità di *tunneling* può essere agilmente calcolata con il metodo WKB descritto in precedenza.

Nel caso in cui esista, in Ω, un potenziale variabile in x, chiamiamolo $U(x)$, tale da richiedere l'abbandono della concezione ad onda piana delle funzioni d'onda, almeno nella loro componente lungo x, è tuttavia ancora possibile

giungere ad una espressione generale per τ_{tun}, a patto di ricorrere all'approssimazione semi–classica. La generica autofunzione confinata ψ_n, con energia $E_x = E_n$, $n \in \mathbb{Z}^+$, può essere espressa come segue (facendo uso dell'Eq.2.6)

$$\phi_n(\mathbf{r}) = \frac{C_n}{\sqrt{k_n}} e^{i\int_{x_0}^{x} k_n(x')\mathrm{d}x'} \cdot \frac{1}{\sqrt{S}} e^{i(\mathbf{k}_\perp \cdot \mathbf{r}_\perp)} \tag{2.19}$$

in cui la funzione di stato è espressa dal prodotto di una componente piana, parallela alla superficie S, e da un a parte che dipende dal potenziale di confinamento lungo x.

$$k_n(x) = \frac{\sqrt{2\mu(E_n - U(x))}}{\hbar}$$

è il vettore d'onda semi–classico. La costante C_n, infine, deriva solamente dall'esigenza di normalizzazione.

Il flusso F_{in} è dato dalla sola componente progressiva dell'onda nella direzione x. In accordo con le ipotesi di semi–classicità, la derivata prima di ψ_n può essere valutata considerando il pre–fattore $\frac{1}{k_n}$

$$F_{in} = \frac{1}{S}\frac{\hbar}{\mu}|C_n|^2$$

con C_n che, secondo le regole di quantizzazione [10], vale

$$C_n = \sqrt{\frac{1}{2\int_{x_0}^{x_1} \frac{1}{k_n(x)\mathrm{d}x}}}$$

dove x_0 e x_1 sono i punti classici di inversione del moto, tali che $U(x_0) = U(x_1) = E_n$. Sostituendo, otteniamo l'espressione per tempo medio di impatto contro la barriera

$$\tau_{att}(E_n) = (J_{in} \cdot S)^{-1} = 2\mu \int_{x_0}^{x_1} \frac{1}{\sqrt{2\mu[E_n - U(x)]}} \tag{2.20}$$

L'inverso di τ_{att} è la cosiddetta *attempt frequency*, *i.e* la frequenza con la quale la particella confinata interagisce con la barriera e "tenta" il *tunneling*. Il tasso di *tunneling*, a sua volta, è dato dall'*attempt frequency* moltiplicata per la trasparenza di barriera, come indica l'Eq.2.14

$$\frac{1}{\tau_{tun,q}(E_n)} = \frac{T(E_n)}{\tau_{att}(E_n)} \tag{2.21}$$

Da notare infine come, avendo utilizzato l'approssimazione semi–classica, si

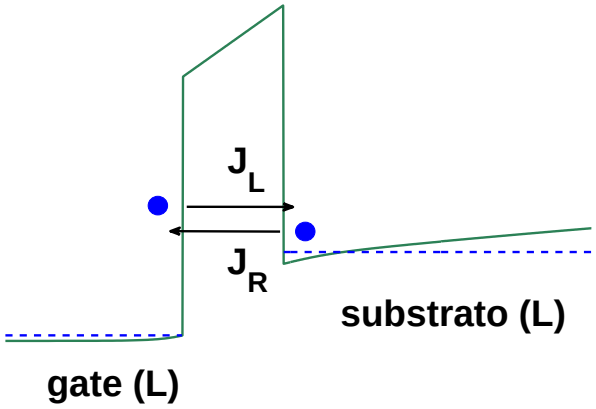


Figura 2.6: *Flussi di carica tra substrato e gate e viceversa. Sono indicati anche i livelli di Fermi nei due elettrodi.*

richieda che il potenziale $U(x)$ non sia estremamente variabile, cosa che nei dispositivi MOS è, tutto sommato, vero.

§2.1.5 Corrente di tunneling

La corrente di *tunneling* che attraversa la barriera di potenziale costituita dall'ossido di *gate* (o di un materiale *High–k*) può essere concepita come differenza tra i flussi di carica che scorrono da una regione di semiconduttore all'altra. Il flusso di portatori proveniente da un elettrodo è la somma dei contributi di tutti gli stati occupati della banda. D'ora in avanti ci riferiremo ad una situazione del tipo rappresentato in Fig. 2.6. Gli elettroni del substrato (R) danno luogo ad una densità di corrente di *tunneling* verso il *gate* (L)

$$J_R = \frac{1}{S} \sum_{\nu} \frac{-q}{\tau_{tun,\nu}} f_R(E_\nu) \left[1 - f_L(E_\nu)\right] \tag{2.22}$$

dove si è indicato con q la carica elementare, con f_R e f_L le distribuzioni di Fermi–Dirac nei due elettrodi e con l'indice ν il relativo stato della banda di conduzione. Come si anticipava, il processo è elastico, *i.e.* l'energia si conserva nel passaggio da un elettrodo all'altro. Ciascun livello energetico ν contribuisce al flusso in ragione del suo tasso $\frac{1}{\tau_{tun,\nu}}$. Un'espressione del tutto analoga all'Eq. 2.22 fornisce la densità di corrente J_L che dal *gate* si riversa sul substrato. La differenza tra le due da la densità di corrente netta.

Per procedere nei calcoli ed ottenere un'espressione compatta, occorre caratterizzare gli stati energetici dei portatori che sono coinvolti nel processo.

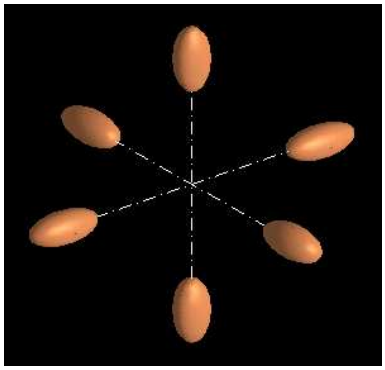


Figura 2.7: *Minimi ellissoidali della banda di conduzione del silicio. Le direzioni corrispondono alla terna (k_x, k_y, k_z), vettori generatori del reticolo reciproco.*

La banda di conduzione del silicio (dal momento che stiamo considerando i soli elettroni) presenta sei minimi alla stessa energia. In condizione di banda piatta, le superfici isoenergia nello spazio reciproco $\mathbf{k}$ sono di forma ellissoidale (vedi Fig.2.7) con i centri posti sulle direzioni dei versori del reticolo reciproco. Analiticamente, la relazione di dispersione può essere quindi così scritta

$$E(\mathbf{k}) = E_c + \frac{\hbar^2(k_x - k_{x,0})^2}{2m_x} + \frac{\hbar^2(k_y - k_{y,0})^2}{2m_y} + \frac{\hbar^2(k_z - k_{z,0})^2}{2m_z} \qquad (2.23)$$

con $\mathbf{k}_0$ che rappresenta la posizione del centro di ogni ellissoide nello spazio reciproco. Nel processo tecnologico CMOS il substrato di silicio viene cresciuto secondo una direzione cristallografica $\langle 100 \rangle$, per usare gli indici di Miller, e questo allo scopo di minimizzare la densità atomica superficiale ovvero i *dangling bonds* all'interfaccia con l'ossido di silicio. Nello spazio del reticolo reciproco, individuabile a partire da una terna cartesiana di vettori (k_x, k_y, k_z), la direzione che corrisponde al momento diretto verso l'interfaccia è proprio k_x, se x è la direzione spaziale ortogonale alla superficie di barriera S. Lungo questa direzione giace l'asse maggiore di due dei sei minimi ellissoidali, che vengono perciò detti minimi *longitudinali*. Gli altri quattro sono disposti sul piano ortogonale alla direzione [100] e presentano, nella direzione k_x, gli assi minori dell'ellissoide. Per questo vengono chiamati minimi *trasversali*. Dal momento che il *tunneling* avviene proprio lungo la direzione (100) nello spazio reale, quindi lungo [100] nello spazio reciproco, gli elettroni esistenti sui due minimi longitudinali manifestano una massa efficace $m_L = 0.98m_0$, quelli sui minimi trasversali invece $m_T = 0.19m_0$. In condizioni di banda piatta, gli stati elettronici del *bulk* possono essere approssimati ad onde piane del tipo nell'Eq. 2.16, per le quali vale una relazione di dispersione quadratico, vedi Eq. 2.23.

Quando, invece, si applica una polarizzazione alla struttura, la banda di

conduzione nel semiconduttore si piega, in una rappresentazione spaziale, e perciò gli stati elettronici cambiano forma. In particolare, se la polarizzazione è positiva al *gate*, nei pressi dell'interfaccia $SiO_2 - Si(bulk)$ il piegamento della banda di conduzione è tale da disegnare una buca di potenziale. L'effetto di un potenziale esterno sugli stati elettronici di un reticolo cristallino può essere trattato efficacemente attraverso il formalismo della *Envelope Function* [9]. In presenza di un potenziale lungo x, esso permette di approssimare le autofunzioni degli elettroni

$$\psi_{n\mathbf{k}}(\mathbf{r}) \approx \xi_n(x) e^{i(k_y y + k_z z)}$$

in cui si può mostrare come l'inviluppo $\xi_n(x)$ è soluzione di una sorta di equazione d'onda

$$-\frac{\hbar^2}{2m_x}\frac{\partial^2}{\partial x^2}\xi_n + E_c(x)\xi_n = \varepsilon_n \xi_n$$

Se il profilo $E_c(x)$ nei pressi dell'interfaccia ossido–semiconduttore è tale da dar luogo ad una quantizzazione dei livelli energetici ε_n, gli stati della banda di conduzione si suddividono in sottobande bidimensionali, caratterizzate dalla relazione di dispersione

$$E_n(\mathbf{k}) = \varepsilon_n + \frac{\hbar^2 k_y^2}{2m_y} + \frac{\hbar^2 k_z^2}{2m_z} \tag{2.24}$$

La Fig. 2.8 mostra l'andamento del fondo della banda di conduzione del substrato sotto l'effetto di una polarizzazione positiva applicata al *gate*: nella regione di canale la banda di conduzione si suddivide in sottobande quantizzate. Per energie elettroniche superiori alla sommità della buca di confinamento, la struttura della banda riacquista la sua forma originale (Eq. 2.23). Gli effetti della quantizzazione diventano sempre più marcati in concomitanza dello *scaling*, in quanto il suo effetto è quello di accentuare la buca di potenziale creata dalla banda di conduzione. Per tensioni di *gate* negative inferiori alla tensione di *flat band*, invece, non sono presenti fenomeni di quantizzazione nel substrato, dal momento che si trova in accumulo di portatori anziché in inversione, mentre è sul polisilicio di *gate* che si riscontreranno dei lievi effetti. Tuttavia essi sono di modesta entità e nel corso della presente trattazione verrano trascurati.

Ai fini del calcolo della corrente di *tunneling*, la divisione in sottobande influenza l'enumerazione degli stati elettronici e sulla forma dei tassi che compaiono nell'Eq. 2.22.

Per gli stati continui, come abbiamo visto, il tempo medio di *tunneling* vale (Eq. 2.18)

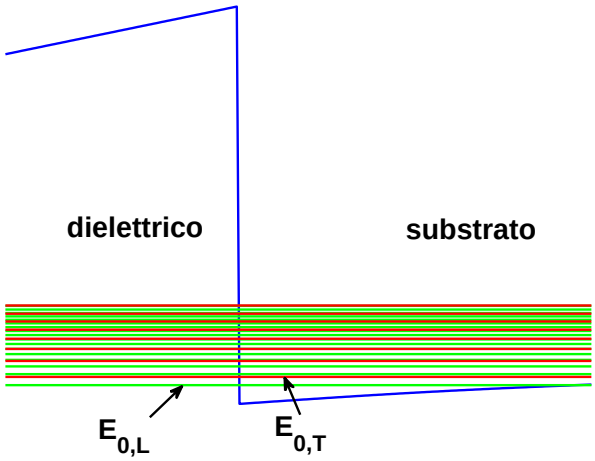


Figura 2.8: *Diagramma a bande di una struttura MOS in cui è applicata una polarizzazione positiva tale da provocare una quantizzazione nella zona di canale. Sono indicati alcuni livelli longitudinali, in verde, e trasversali, in rosso.*

$$\frac{1}{\tau_{tun,v,\mathbf{k}}} = T(E_x)\frac{\hbar k_x}{m_{x,v}}\frac{1}{L}$$

dove è indicato con v la valle in cui risiedono gli elettroni, mentre il vettore d'onda $\mathbf{k} = (k_x, k_y, k_z)$ enumera gli stati nell'intorno del minimo della banda in questione. T è come sempre la trasparenza di barriera, k_x ed $m_{x,v}$ rappresentano la componente del vettore d'onda nella direzione di *tunneling* e la massa efficace nella medesima direzione, rispettivamente. Infine L è la lungezza del *bulk*.

Per gli stati delle sottobande quantizzate, invece, il calcolo del tempo medio di *tunneling* risulta, secondo l'approssimazione semi–classica WKB vista in precedenza (Eq. 2.21)

$$\frac{1}{\tau_{tun,v,n,\mathbf{k}_\perp}} = \frac{T(\varepsilon_n)}{\tau_{att,v}(\varepsilon_n)} = \frac{T(\varepsilon_n)}{\int_{x_0}^{x_1}\sqrt{\frac{2m_{x,v}}{\varepsilon_n - E_c(x)}}\mathrm{d}x}$$

dove anche in questo caso la massa efficace nella direzione del moto $m_{x,v}$ dipende dalla valle v che stiamo considerando. Il totale degli stati elettronici è questa volta enumerato con l'indice di sottobanda n e dal vettore d'onda bidimensionale trasverso $\mathbf{k}_\perp = (k_y, k_z)$.

Data la simmetria del sistema, assumiamo che la trasparenza di barrieera

dipenda solamente dalla componente energetica nella direzione del moto. Per gli stati continui essa vale $E_x = E_c + \frac{\hbar^2 k_x^2}{2m_x}$, mentre per quelli quantizzati ε_n. Come massa di *tunneling* dobbiamo considerare la massa efficace nella direzione del moto all'interno del mezzo attraversato, *i.e.* l'ossido di silicio o il materiale *High–k* eventuale. Essa, chiamata anche massa efficace di *tunneling*, dipende oltre che dal reticolo cristallino del mezzo attraversato, anche dall'energia della particella viaggiante. Un opportuno valore costante permette tuttavia di approssimare con buona accuratezza la caratteristica di *tunneling* [11].

Stabilita l'espressione per i tassi di *tunneling* e la disposizione energetica degli stati elettronici, la formula generale nell'Eq. 2.22 può essere descritta per ogni valle, separando il contributo degli stati continui da quelli quantizzati

$$J_{R,v} = -\frac{2q}{S}\sum_n \frac{T(\varepsilon_n)}{\tau_{att,v}(\varepsilon_n)}\sum_{k_y}\sum_{k_z} f_R(E)[1-f_L(E)] - \\ - \frac{2q}{SL}\sum_{k_x} T(E_x)\frac{\hbar k_x}{m_{x,v}}\sum_{k_y}\sum_{k_z} f_R(E)[1-f_L(E)]$$

Il primo addendo è associato agli stati delle sottobande quantizzate, il secondo invece agli stati continui al di sopra di essi. Il fattore 2 rende conto della molteplicità dello *spin*. Passando dalle sommatorie agli integrali per poter procedere analiticamente, per i vettori d'onda quasi–continui i termini S ed L si elidono nel momento in cui vado a considerare la densità di stati nello spazio $\mathbf{k}$, che emerge dall'introduzione dei fattori differenziali.

$$J_{R,v} = -\frac{q}{2\pi^2}\sum_n \frac{T(\varepsilon_n)}{\tau_{att,v}(\varepsilon_n)}\iint_{k_y,k_z} f_R(E)[1-f_L(E)]\mathrm{d}k_y\mathrm{d}k_z - \\ - \frac{q}{4\pi^3}\int_{k_x} T(E_x)\frac{\hbar k_x}{m_{x,v}}\left(\iint_{k_y,k_z} f_R(E)[1-f_L(E)]\mathrm{d}k_y\mathrm{d}k_z\right)\mathrm{d}k_x$$

Se ora si sottrae il flusso complementare $J_{v,L}$ e si svolgono gli integrali nella variabile energia, ottenibile in funzione dei vettori d'onda dalle Eqs. 2.23, 2.24, si ottiene un'espressione per la corrente netta di *tunneling* dovuta alla valle v.

Gli eventuali stati quantizzati presenti nella buca di potenziale nel substrato generano una corrente netta

$$J_{v,Quant.} = \frac{4q\pi m_{SD,v}kT}{\hbar^2}\sum_n \frac{T_v(\varepsilon_n)}{\tau_{att,v}(\varepsilon_n)}\ln\left(\frac{1+e^{-\frac{\varepsilon_n - E_{f,R}}{kT}}}{1+e^{-\frac{\varepsilon_n - E_{f,L}}{kT}}}\right) \tag{2.25}$$

Il fattore $m_{SD,v}$ è detto massa efficace per il calcolo della densità di stati e deriva dal cambio di variabile $\mathbf{k} \rightarrow E$. Il suo valore dipende dalla valle considerata: per le longitudinali si ha $m_{SD,v} = m_T$, per le trasversali $m_{SD,v} = \sqrt{m_T \cdot m_L}$. Il rapporto contenuto nel logaritmo rende nulla la corrente all'equilibrio termodinamico, *i.e.* quando i livelli di Fermi nelle due regioni sono allineati.

Alla corrente degli stati quantizzati si aggiunge quella degli stati continui

$$J_{v,Cont.} = \frac{4q\pi m_{SD,v} kT}{\hbar^3} \int_{E_0}^{+\infty} T(E_x) \ln \left(\frac{1 + e^{-\frac{E_x - E_{f,R}}{kT}}}{1 + e^{-\frac{E_x - E_{f,L}}{kT}}} \right) \mathrm{d}E_x \tag{2.26}$$

con l'integrazione che inizia dall'energia E_0, limite inferiore dell'energia degli stati continui. Essa dipende dunque dalla polarizzazione applicata: se la polarizzazione al *gate* non è tale da generare livelli quantizzati nel substrato, E_0 va posta pari al massimo tra $E_c(-t_{ox})$ e $E_c(0)$.

In definitiva, la densità di corrente imputabile alla singola valle v

$$J_v = J_{v,Quant.} + J_{v,Cont.}$$

mentre considerando il contributo collettivo di tutte le valli

$$J_{tun} = 2J_L + 4J_T$$

dove J_L e J_T sono le densità dovute rispettivamente alle valli longitudinali e alle trasversali.

Spendiamo infine qualche parola anche per le lacune in banda di valenza. Per il calcolo del loro contributo alla densità di corrente valgono i medesimi ragionamenti visti sinora. L'unica differenza di rilievo riguarda la disposizione degli stati nella banda di valenza. Essa infatti ha 2 soli minimi, leggermente scostati in energia per effetti di accoppiamento *spin*–orbita (fenomeno quantistico relativistico). Questi minimi hanno forma sferica, nello spazio $\mathbf{k}$, ma hanno curvatura diversa: si dividono in lacune pesanti m_H, e lacune leggere m_L.

§2.2 Corrente assistita da 1 difetto (1TAT)

In molti casi, il modello di *tunneling* diretto/Fowler–Nordheim non è in grado di spiegare l'andamento sperimentale delle correnti di *leakage*, specialmente

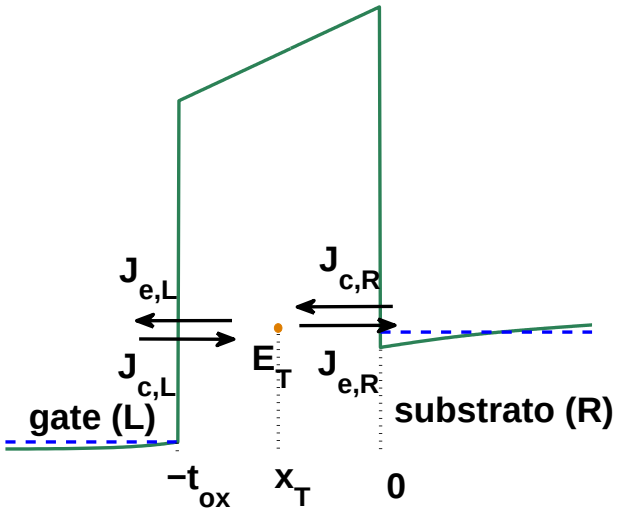


Figura 2.9: *Flussi di cattura e di emissione da e verso lo stato trappola in x_T di energia E_T. Sono indicati a tratteggio i livelli di Fermi degli elettrodi di gate e substrato.*

per basse polarizzazioni di *gate* [12]. Essa infatti aumenta in modo sensibile soprattuto a seguito di uno stress elettrico subito dal dielettrico. Questo surplus di corrente viene chiamato *Stress Induced Leakage Current* (SILC), e la sua origine è da ricercare nel fatto che il flusso di carica che attraversa il dielettrico alla lunga interagisce col reticolo cristallino, generando localmente dei difetti reticolari. Questi stati "trappola", localizzati spazialmente e dall'energia compresa nel *gap* proibito, fungono da "ponte" per il passaggio di ulteriore carica elettrica. Essendo un meccanismo che può essere molto efficiente [13], è necessario sviluppare un modello che sia in grado di spiegarlo, per poterne estrarre informazioni significative riguardo alla potenza dissipata e alla carica trasportata. Il meccanismo di trasporto viene comunemente denominato *Trap Assisted Tunneling* (TAT), ed in questa sezione ne verrà analizzato il caso ad una singola trappola.

§2.2.1 Modello 1TAT

Per descrivere quantitativamente il 1TAT è possibile adottare un modello unidimensionale che descrive il passaggio di carica attraverso lo strato dielettrico come un processo a due passi: cattura di un portatore da un elettrodo da parte della trappola e successivo rilascio all'altro elettrodo [14]. Benché siano possibili processi di cattura/emissione che coinvolgono anche le lacune, in questa trattazione si prenderanno in considerazione solo gli elettroni, che tipicamente danno il contributo dominante. la densità di corrente associata

alla cattura di portatori da un elettrodo da parte di uno strato di trappole non interagenti è esprimibile come

$$J_c = qc(1 - f_T)N_{s,T} \tag{2.27}$$

dove q è la carica elementare, $c = \tau_c^{-1}$ è il tasso di cattura dall'elettrodo, f_T è la probabilità di occupazione dello stato–trappola. Il termine $(1 - f_T)$ deriva dal principio di esclusione: la transizione di cattura avviene solamente se lo stato trappola è vuoto. $N_{s,T}$ è la densità di trappole per unita di superficie, quindi il prodotto $(1 - f_T)N_{s,T}$ rappresenta il numero medio di trappole libere per unità di area. Allo stesso modo, la corrente originata dall'emissione da parte delle trappole verso l'elettrodo è

$$J_e = qef_T N_{s,T} \tag{2.28}$$

dove $e = \tau_e^{-1}$ è il tasso di emissione verso l'elettrodo. La corrente netta tra strato di trappole e elettrodo è data dalla differenza tra i flussi di cattura e di emissione, sia per il substrato (R) che per il *gate* (L), come mostrato in Fig. 2.9. In condizione stazionarie, i flussi netti di carica scambiati dalle trappole con gli elettrodi sono uguali a zero

$$J_{1TAT} = (J_{c,L} - J_{e,L}) = (J_{c,R} - J_{e,R}) = 0 \tag{2.29}$$

Questa relazione permette di valutare la corrente di 1TAT in condizione di polarizzazione, tuttavia bisogna essere in grado di calcolare in modo agevole i tassi di cattura e di emissione, che compaiono nelle realazioni costitutive nelle Eqs. 2.27 e 2.28. In letteratura esistono vari approcci: alcuni si basano su relazioni di tipo empirico sperimentale [15], altri partendo da principi primi [16], altri ancora con modelli statistici [19][20]. Nel presente lavoro di tesi si dà invece una rappresentazione precisa degli stati elettronici negli elettrodi, rinunciando però ad una descrizione miscroscopica delle trappole [17][18], interpretate secondo lo stile della teoria della generazione/ricombinazione SHR [21][22].

§2.2.2 Calcolo dei tassi di cattura

Si consideri un sistema costituito da una regione di semiconduttore, separata dal sito trappola da uno strato di isolante. Il calcolo dei tassi di cattura avviene in maniera analoga a quanto fatto in precedenza per le correnti di *tunneling*. Esso è dato dalla somma dei contributi elettronici di tutti gli stati della banda di conduzione

$$c = \sum_{\nu} \sigma_{c,\nu} F_{\nu} f(E_{\nu}) \tag{2.30}$$

dove il pedice ν enumera gli stati della banda dell'elettrodo e $F_{c,\nu}$ rappresenta il flusso di probabilità originato dallo stato ν che raggiunge la trappola. Questo flusso viene "catturato" dalla trappola in ragione della sua *cross section*, sezione efficace di interazione $\sigma_{c,\nu}$. Il prodotto $\sigma_{c,\nu}F_{\nu}$ rappresenta dunque la probabilità che, nell'unità di tempo, avvenga la cattura dell'elettrone che occupa lo stato ν. $f(E_{\nu})$ è infine la probabilià di occupazione del suddetto stato, secondo la statistica di Fermi–Dirac.

La *cross section* caratterizza fisicamente le proprietà del difetto e le proprietà di interazione con gli elettroni dell'elettrodo. Geometricamente può essere interpretata come un'"area di cattura", e infatti ha le dimensioni di una superficie. Il suo valore, da determinarsi sperimentalmente, si suppone in generale dipendente dall'energia della trappola e da quella dello stato di partenza

$$\sigma_{c,\nu} = \sigma_c(E_{\nu}, E_T)$$

Si danno, comunemente, due casi estremi di questa dipendenza: *cross section* costante con l'energia del portatore

$$\sigma_c(E_{\nu}, E_T) = \begin{cases} \sigma_0 & \text{per } E_{\nu} > E_T, \\ 0 & \text{per } E_{\nu} < E_T. \end{cases} \tag{2.31}$$

oppure di tipo deltiforme

$$\sigma_c(E_{\nu}, E_T) = s_0 \cdot \delta(E_{\nu} - E_T) \tag{2.32}$$

L'Eq. 2.32 modellizza una cattura completamente anelastica: il portatore che raggiunge la trappola con energia $E_{\nu} > E_T$ ha sempre una probabilità non nulla di essere catturato. Una volta avvenuta la cattura la particella va ad occupare lo stato trappola di energia E_T e l'energia in eccesso viene ceduta al reticolo sottoforma di moti vibrazionali, *i.e.* fononi. Evidenze sperimentali [23][17] mostrano come questa sembri essere la modellizzazione più vicina alla realtà per i dispositivi elettronci di nostro interesse.
L'Eq. 2.32 modellizza una cattura che può avere luogo solo se il portatore ha la stessa energia della trappola, dando vita ad un processo perfettamente elastico.
Si possono poi immaginare tutti i casi intermedi, in cui la forma funzionale

della sezione di cattura possa assumere le più svariate forme, a seconda di come la trappola interagisce col reticolo e delle sue proprietà fisiche [24].

Il flusso di probabilità che raggiunge la trappola può essere espresso come

$$F_{c,\nu} = T(x_T, E_\nu) F_\nu$$

dove $T(x_T, E_\nu)$ è la trasparenza di barriera di potenziale che separa la trappola dall'elettrodo, mentre F_ν è il flusso di probabilità che "sbatte" contro la barriera stessa. Come si è visto in precedenza, il flusso incidente sulla barriera ha espressione dipendente dalla forma dello stato ν, sia esso continuo o quantizzato. Ancora, per ciascuna valle v della banda di conduzione, si ha

$$F_{c,v}(\mathbf{k}) = T(x_T, E_x) \frac{\hbar k_x(E_x)}{m_{x,v}} \frac{1}{L} \tag{2.33}$$

per gli stati continui, mentre per quelli quantizzati

$$F_{c,v}(\varepsilon_n, \mathbf{k}) = \frac{T(x_T, \varepsilon_n)}{\tau_{att,v}(\varepsilon_n)} \tag{2.34}$$

Ragionando come nella precedente sezione, si separano i termini relativi agli stati continui e a quelli quantizzati. Svolgendo gli integrali e mettendo in luce la variabile energia, si ottiene

$$c_{v,Quant.} = \frac{4\pi m_{SD,v}}{\hbar^2} \sum_n \frac{T(x_T, \varepsilon_n)}{\tau_{att,v}(\varepsilon_n)} \int_{\varepsilon_n}^{+\infty} \sigma_c(\varepsilon_n, E_\perp, E_T) f(E) \mathrm{d}E_\perp \tag{2.35}$$

per gli stati quantizzati, mentre per quelli continui

$$c_{v,Cont.} = \frac{4\pi m_{SD,v}}{\hbar^3} \int_{E_0}^{+\infty} T(x_T, E_x) \left(\int_{E_x}^{+\infty} \sigma_c(E_x, E_\perp, E_T) f(E) \mathrm{d}E_\perp \right) \mathrm{d}E_x \tag{2.36}$$

In assenza di quantizzazione, $c_{v,Quant.} = 0$ ed E_0 rappresenta il fondo della banda di conduzione all'interfaccia ossido–semiconduttore. In presenza di livelli quantizzati, E_0 va posta pari alla sommità della buca di potenziale di confinamento.

Il tasso di cattura associato alla singola valle v è dato dalla somma dei due contributi

$$c_v = c_{v,Quant.} + c_{v,Cont.}$$

mentre il tasso globale è dato dalla somma dei contriburi delle singole valli longitudinali (L) e trasversali (L)

$$c = 2c_L + 4c_T$$

§2.2.3 Calcolo dei tassi d'emissione

Per quanto riguarda il tasso di emissione da un sito trappola ad un elettrodo, si può dare un'espressione generale formalmente analoga all'Eq. 2.30

$$e = \sum_{\nu} \sigma_{e,\nu} F_{e,\nu} [1 - f(E_\nu)] \tag{2.37}$$

in cui il prodotto $\sigma_{e,\nu} F_{e,\nu}$ indica il tasso con cui avviene l'emissione del portatore dal sito trappola verso lo stato ν dell'elettrodo. Al flusso emesso si può dare la stessa forma funzionale del corrispondente flusso di cattura

$$F_{e,\nu} = F_{c,\nu}$$

Il coefficiente $\sigma_{e,\nu}$ può essere determinato attraverso argomenti termodinamici [21][22]. In condizioni di equilibrio termodinamico, infatti, i flussi di cattura e di emissione originati dall'interazione della trappola con l'elettrodo si bilanciano. Per ciascuno stato ν quindi vale

$$\sigma_{e,\nu} F_{e,\nu} [f(E_\nu)] f_T = \sigma_{c,\nu} F_{c,\nu} f(E_\nu)(1 - f_T) \tag{2.38}$$

con l'accorgimento che all'equilibrio termodinamico la funzione di occupazione della trappola è regolata anch'essa dalla statistica di Fermi–Dirac, uniforme in tutto il sistema: $f_T = f(E_T)$. Utilizzando questa considerazione nell'Eq. 2.38 si ottiene il legame tra le *cross section* di cattura e di emissione

$$\sigma_{e,\nu} = \sigma_{c,\nu} \exp\left(-\frac{E_\nu - E_f}{kT}\right) \tag{2.39}$$

che se sfruttata nell'Eq. 2.37 può dare un'espressione molto significativa per il tasso di emissione

$$\frac{1}{\tau_e} = \frac{1}{\tau_c} \exp\left(-\frac{E_f - E_T}{kT}\right) \tag{2.40}$$

da cui si evince come l'emissione domini sulla cattura quanto più l'energia

del sito trappola è alta rispetto al livello di fermi dell'elettrodo. Il termine esponenziale può invece essere interpretato come la probabilità che il portatore intrappolato all'energia E_T raggiunga, per agitazione termica, il livello di Fermi del semiconduttore. L'Eq. 2.40 permette inoltre di esprimere le correnti nette ai due elettrodi come segue

$$\begin{aligned} J_L &= J_{c,L} - J_{e,L} = qN_{s,T}c_L \left[1 - \tfrac{f_T}{f_L(E_T)}\right] \\ J_R &= J_{c,R} - J_{e,R} = qN_{s,T}c_R \left[1 - \tfrac{f_T}{f_R(E_T)}\right] \end{aligned} \tag{2.41}$$

dove appare evidente che il flusso di cattura domini su quello di emissione nel momento in cui la trappola è non–occupata con maggior probabilità rispetto all'elettrodo considerato.

§2.2.4 Funzione di occupazione e corrente 1TAT

Una volta calcolati i tassi di cattura e di emissione, non resta che trovare la funzione di occupazione delle trappole, f_T, per individuare finalmente la corrente di 1TAT. Usando l'Eq. 2.41 e imponendo che, in presenza di polarizzazione, non si manifestino fenomeni di accumulo/rilascio di carica elettrica all'interno delle trappole stesse in funzione del tempo, ovvero che le correnti siano uguali

$$J_{c,L} - J_{e.L} = J_{e,R} - J_{c,R}$$

è possibile stabilire f_T

$$\frac{1}{f_T} = \frac{1}{f_L(E_T)}\frac{c_L}{c_L + c_R} + \frac{1}{f_R(E_T)}\frac{c_R}{c_L + c_R} \tag{2.42}$$

che è molto significativa, dacché mostra come l'occupazione delle trappole sia una somma pesata delle funzioni di Fermi ai due elettrodi. Come si vede, la f_T tende alla funzione di Fermi dell'elettrodo a cui corrisponde il tasso di cattura più elevato. I tassi medi di cattura sono a loro volta dipendenti dalla probabilità di *tunneling* attraverso la barriera che separa il sito trappola dall'elettrodo: ne consegue che la f_T è più simile alla distribuzione di Fermi dell'elettrodo più vicino.

Nota la f_T, si può ora dare una forma esplicita alla corrente di 1TAT stazionaria, sostituendo l'Eq. 2.42 nell'Eq. 2.41

$$J_{1TAT} = qN_{s,T}\left[\frac{f_L(E_T)}{c_L} + \frac{f_R(E_T)}{c_R}\right]^{-1}\left[f_L(E_T) - f_R(E_T)\right] \tag{2.43}$$

È interessante rielaborare, nell'Eq. 2.43, il fattore dipendente dai tassi di cattura

$$\frac{c_L}{f_L(E_T)} = c_L + e_L, \qquad \frac{c_R}{f_R(E_T)} = c_R + e_R$$

per mostrare come, dopo una breve analisi a partire da questa considerazione, la corrente di 1TAT sia in un certo senso massimizzata dalle trappole i cui tassi di cattura e di emissione si eguagliano.

Si è visto come calcolare la corrente di 1TAT relativa ad un singolo strato di difetti, localizzati quindi sia nella posizione spaziale x_T che energetica E_T. Il discorso può essere tuttavia facilmente generalizzato ad una distribuzione di trappole qualunque $N_T(x_T, E_T)$, ovviamente sempre non interagenti. È sufficiente sostituire la densità per unità di superficie $N_{s,T}$ con $N_T \mathrm{d}x_T \mathrm{d}E_T$ e integrare opportunamente. Definendo la densità di corrente normalizzata:

$$j(x_T, E_T) = q \left[\frac{f_L(E_T)}{c_L(x_T, E_T)} + \frac{f_R(E_T)}{c_R(x_T, E_T)} \right]^{-1} [f_L(E_T) - f_R(E_T)]$$

la corrente totale non è altro che

$$J_{1TAT} = \iint N_T(x_T, E_T) j(x_T, E_T) \mathrm{d}x_T \mathrm{d}E_T \tag{2.44}$$

§2.3 Corrente assistita da 2 difetti (2TAT)

Nonostante il meccanismo di conduzione 1TAT sia in grado di spiegare molti dei casi di conduzione anomala manifestati da dispositivi a struttura MOS, esistono casi in cui, per nessun valore dei parametri del modello, si è riusciti ad ottenere un accordo accettabile con i dati sperimentali sul SILC [25]. Una delle possibili spiegazioni è che sinora si sono considerati i difetti indipendenti tra loro, ovvero che non sono contemplati fenomeni di emissione/cattura tra le trappole stesse. Questa approssimazione, tenuta implicita nella precedente sezione, può non essere più valida nel momento in cui si sia in presenza di densità di difetti molto elevate, anche solo localmente, tale da consentire una loro possibile interazione.

Anche in questo caso sono presenti in letteratura numerosi lavori, ognuno dei quali elabora il modello da un proprio punto di vista. Da chi si basa su considerazioni puramente statistiche [26], a chi invoca teorie percolative [27], a

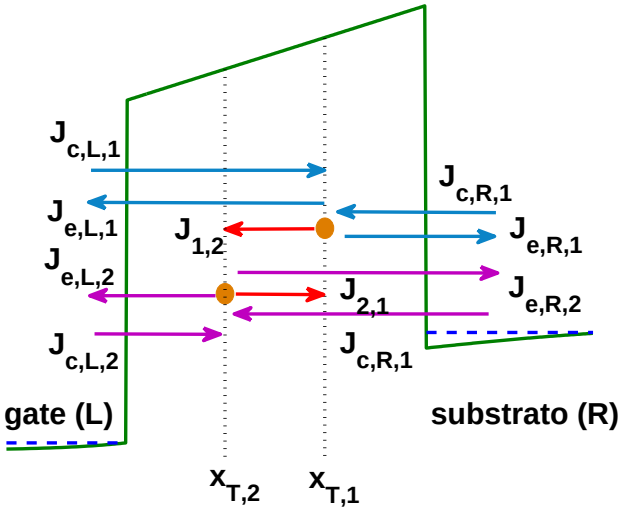


Figura 2.10: *Rappresentazione schematica dei flussi di cattura e di emissione tra le trapole e gli elettrodi e tra le trappole medesime. Sono indicati a tratteggio i livelli di Fermi degli elettrodi di gate e substrato.*

chi infine generalizza il modello studiato nella precedente sezione [29][30], attenendosi ad una descrizione sullo stile cattura/emissione alla SHR [21]. Per il presente lavoro verrà adottato quest'ultimo modello, poiché conserva un certo rigore analitico (quando tutti gli altri ricorrono a significative approssimazioni) e allo stesso tempo non diverge in termini di complessità. Dopo una descrizione puntuale dal punto di vista formale, verranno presentate delle simulazioni numeriche ed il loro *fitting* di alcuni dati sperimentali, a riprova dell'efficacia di questo modello.

§2.3.1 Calcolo dei flussi

Consideriamo la situazione rappresentata schematicamente in Fig. 2.10. Rispetto alla conduzione 1TAT, ora ci troviamo nella situazione in cui due trappole, o meglio due strati di trappole, dal momento che siamo in una realtà monodimensionale, cooperano tra loro. Accanto quindi ai soliti flussi di cattura/emissione tra ciascuna trappola e ciascun elettrodo, abbiamo anche una corrente di "scambio" tra i due difetti. Essa ha importanti ripercussioni nel momento di andare a considerare le loro funzioni di occupazione, che saranno quindi interdipendenti. Per analizzare il meccanismo, procediamo come segue.

Consideriamo innanzi tutto l'interazione delle trappole coi soli elettrodi. Le espressioni delle correnti nette verso il substrato e verso il *gate* sono uguali, per ragioni termodinamiche, a quelle ricavate per le trappole singole (Eq. 2.41)

$$\begin{cases} J_{1,L} = J_{1,c,L} - J_{1,e,L} = qN_{1,T}c_{1,L}\left[1 - \frac{f_{1,T}}{f_L(E_{1,T})}\right] \\ J_{1,R} = J_{1,c,R} - J_{1,e,R} = qN_{1,T}c_{1,R}\left[1 - \frac{f_{1,T}}{f_R(E_{1,T})}\right] \end{cases} \tag{2.45}$$

$$\begin{cases} J_{2,L} = J_{2,c,L} - J_{2,e,L} = qN_{2,T}c_{2,L}\left[1 - \frac{f_{2,T}}{f_L(E_{2,T})}\right] \\ J_{2,R} = J_{2,c,R} - J_{2,e,R} = qN_{2,T}c_{2,R}\left[1 - \frac{f_{2,T}}{f_R(E_{2,T})}\right] \end{cases} \tag{2.46}$$

dove i pedici 1 e 2 sono relativi ai due strati di trappole.

Il termine di interazione tra di esse invece può essere scritto così

$$\begin{cases} J_{12} = qN_{2,T} \cdot (1 - f_{2,T}) \cdot c_{21} \\ J_{21} = qN_{1,T} \cdot (1 - f_{1,T}) \cdot c_{12} \end{cases} \tag{2.47}$$

dove con J_{12} si intende la corrente generata dagli elettroni che fluiscono dalla prima trappola alla seconda, proporzionale al numero di stati liberi disponibili all'arrivo dati dal prodotto $N_{2,T}(1-f_{2,T})$, e dal termine c_{21} che indica il tasso di cattura da parte della seconda trappola di un elettrone proveniente dalla prima. Considerazioni analoghe per valgono per J_{21}. Come si nota, si è preferito mettere in luce soltanto i tassi di cattura e non quelli di emissione, dal momento che la cattura da parte della seconda trappola di un portatore dalla prima è analoga all'emissione di un portatore dalla prima e diretta verso la seconda. Le due scritture sono cioè sostanzialmente equivalenti.

§2.3.2 Tassi di cattura inter–trappola

I tassi di cattura e di emissione da e verso gli elettrodi sono gli stessi ricavati per il caso 1TAT. Quello che bisogna definire esplicitamente, ora, sono i tassi di cattura inter–trappola

$$\begin{cases} c_{21} = N_{1,T} \cdot f_{1,T} \cdot \sigma_2 \cdot \frac{T_{12}}{\tau 1} \\ c_{12} = N_{2,T} \cdot f_{2,T} \cdot \sigma_1 \cdot \frac{T_{21}}{\tau 2} \end{cases} \tag{2.48}$$

Prendendo come esempio la prima equazione, essi vengono modellizzati come il prodotto tra il numero di trappole di partenza occupate $N_{1,T} \cdot f_{1,T}$, la *cross section*[1] della trappola di arrivo σ_2 ed un termine che tiene conto del flusso

[1] Le *cross section* d'interazione tra le trappole vengono considerate del tipo completamente inelastico, per semplicità di trattazione e ragionevole approssimazione, vedi Eq. 2.31

quantistico in uscita dalla trappola di partenza $\frac{T_{12}}{\tau 1}$ concepito, come al solito, come una trasparenza di barriera su un tempo di *attempt* caratteristico, che questa volta è considerato un parametro di *fitting* dal momento che non abbiamo informazioni circa le caratteristiche degli autostati degli elettroni confinati nelle trappole. Lo stesso ragionamento è applicabile alla seconda equazione.

Poiché all'equilibrio termodinamico $J_{12} = J_{21}$ e le f_T sono le funzioni di occupazione di Fermi–Dirac, sostituendo l'espressione dei tassi di cattura nell'Eq. 2.48 nell'Eq. 2.47 e uguagliando i termini, si ottiene che le *cross section* sono legate dalla relazione

$$\sigma_2 \frac{T_{21}}{\tau 1} = \sigma_1 \frac{T_{12}}{\tau 2} \cdot e^{\frac{E_{1,T} - E_{2,T}}{kT}}$$

Se definiamo delle *cross section* efficaci in modo tale che

$$\begin{cases} \sigma_1' = \sigma_1 \frac{T_{12}}{\tau 2} \\ \\ \sigma_2' = \sigma_2 \frac{T_{21}}{\tau 1} \end{cases}$$

la relazione diventa, più sinteticamente

$$\sigma_2' = \sigma_1' \cdot e^{\frac{E_{1,T} - E_{2,T}}{kT}} \tag{2.49}$$

Come ultima considerazione sui tassi di cattura, esplicitiamo anche il termine relativo alla trasparenza di barriera T_{12} (non sarà necessario calcolare anche T_{21} dato che l'Eq. 2.49 ne fornisce già un legame[2]) che è scrivibile come una media pesata della trasparenza di barriera, per tenere conto del fatto che l'energia del portatore confinato non è necessariamente quella dello stato (legato) fondamentale, può avere una componente derivante dall'agitazione termica che lo eleva energeticamente.

$$T_{12} = \int_{E_i}^{+\infty} T_{12}(E) p(E) \mathrm{d}E \tag{2.50}$$

in cui E_i corrisponde alla più elevata delle energie tra la prima e la seconda trappola, al di sotto del quale non ho possibilità alcuna di transizione, e $p(E)$ che corrisponde alla funzione peso. Un esempio per essa è una funzione alla Maxwell–Boltzmann

$$p(E) = e^{-\frac{(E - E_j)}{kT}} \frac{1}{kT}$$

[2]L'affermazione è vera se considero i tempi caratteristici $\tau_1 = \tau_2$ e le *cross section* $\sigma_1 = \sigma_2$, in linea di principio non necessariamente vero ma comunque accettabile nella maggior parte dei casi.

dove il pedice j identifica la trappola di partenza.

Sostituendo l'espressione dei tassi delle Eqs. 2.48 nell'Eq. 2.47 siamo in grado di scrivere un'espressione per il flusso di carica netto tra i due strati di trappole

$$J_{TT} = J_{12} - J_{21} = N_{1,T}N_{2,T} \cdot [\sigma'_2 \cdot f_{1,T}(1 - f_{2,T}) - \sigma'_1 \cdot f_{2,T}(1 - f_{1,T})] \quad (2.51)$$

che come si vede hanno dei termini in cui compaiono, moltiplicate tra loro, le rispettive funzioni di occupazione.

§2.3.3 Calcolo delle funzioni di occupazione e corrente 2TAT

Per giungere ad un'espressione ben definita della corrente, occorre infine calcolare le funzioni di occupazione delle trappole $f_{1,T}$ e $f_{2,T}$. Per fare ciò, adottiamo un procedimento analogo a quanto visto per il caso 1TAT: poiché assumiamo di non avere fenomeni di accumulo/rilascio di carica nell'ossido al variare del tempo, la corrente entrante nel dielettrico uguaglia perfettamente quella di uscita, nonché quella calcolata in una sezione trasversale ad una profondità qualsiasi dell'isolante. Prendiamo quindi come riferimenti la densità di corrente al substrato, al *gate* e ad una posizione tale da trovarsi in mezzo ai due strati di trappole, indicandole rispettivamente con J_R, J_L e J_M

$$\begin{cases} J_R = J_{1,R} + J_{2,R} \\ J_M = J_{1,L} + J_{2,R} + J_{TT} \\ J_L = J_{1,L} + J_{2,L} \end{cases} \quad (2.52)$$

per le cui espressioni esplicite è sufficiente effettuare la sostituzione dei singoli contributi di corrente riportati nelle Eqs. 2.45, 2.46, 2.51.

Uguagliandole a due a due otteniamo un sistema di due equazioni in due incognite, $f_{1,T}$ e $f_{2,T}$

$$\begin{cases} J_R(f_{1,T}, f_{2,T}) = J_L(f_{1,T}, f_{2,T}) \\ J_R(f_{1,T}, f_{2,T}) = J_M(f_{1,T}, f_{2,T}, f_{1,T} \cdot f_{2,T}) \end{cases} \quad (2.53)$$

in cui è mostrato come J_M dipenda anche dal prodotto delle due funzioni di occupazione, rendendo il sistema di secondo grado. Svolgendo i calcoli, si ottengono le complicate espressioni analitiche

$$\begin{cases} f_{1,T} = A + Bf_{2,T} \\ f_{2,T}^2 BD + f_{2,T}(AD - BF + C) - (E + FA) = 0 \end{cases} \quad (2.54)$$

con

$$A = \frac{c_{1,R}+c_{1,L}}{\frac{c_{1,R}}{f_R(E_{1,T})}+\frac{c_{1,L}}{f_L(E_{1,T})}} + \frac{N_{2,T}}{N_{1,T}} \frac{c_{2,R}+c_{2,L}}{\frac{c_{1,R}}{f_R(E_{1,T})}+\frac{c_{1,L}}{f_L(E_{1,T})}}$$

$$B = -\frac{N_{2,T}}{N_{1,T}} \frac{\frac{c_{2,R}}{f_R(E_{2,T})}+\frac{c_{2,L}}{f_R(E_{2,T})}}{\frac{c_{1,R}}{f_R(E_{1,T})}+\frac{c_{1,L}}{f_L(E_{1,T})}}$$

$$C = \frac{c_{2,R}}{f_R(E_{2,T})} + \frac{c_{2,L}}{f_R(E_{2,T})} + N_{1,T}\sigma_1'$$

$$D = N_{1,T}(\sigma_2' - \sigma_1')$$

$$E = c_{2,L} + c_{2,R}$$

$$F = N_{1,T}\sigma_2'$$

che ci consentono di risolvere analiticamente il problema. Una volta calcolati, quindi, i valori delle funzioni di occupazione, per ottenere il valore della corrente 2TAT è sufficiente sostituire tali risultati in una qualsiasi delle espressioni della corrente (Eq. 2.52). È interessante constatare come esse dipendano non solo dai tassi di cattura e dai livelli di Fermi degli elettrodi, ma anche dalle densità di trappole stesse: più le densità diminuiscono, più le trappole si comportano come se fossero indipendenti, viceversa più le densità aumentano, più sono interagenti. In sostanza la corrente di perdita dipende dalle densità sia in modo esplicito che in modo implicito, attraverso le f_T. In Fig. 2.11 è mostrato un esempio di corrente 2TAT messo a confronto con le correnti 1TAT che invece si avrebbero se le due trappole agissero singolarmente. È evidente come la loro "collaborazione" aumenti il flusso di carica traportato, rendendo il meccanismo di conduzione 2TAT più efficiente rispetto al 1TAT. Il fatto che questo fenomeno sia più evidente nella parte destra del grafico deriva dai valori dei parametri spaziali ed energetici che caratterizzano le trappole in questo specifico esempio. Per tali valori di tensione, infatti, le trappole interagiscono in maniera significativa, mentre per valori più bassi l'interazione è meno marcata e il contributo di una delle due trappole è dominante rispetto all'altro, riducendo così il 2TAT ad una sorta di doppio 1TAT.

Come considerazione finale, è necesario far notare che il modello appena descritto vale per due strati di trappole la cui densità è di tipo deltiforme nello spazio e nell'energia. Se ci si trovasse nel caso generale in cui $N_T = N_T(x, E)$ e si volesse considerare l'interazione di tutte le trappole, prese a due a due, ci si troverebbe a dover risolvere un sistema non lineare accoppiato il cui numero di equazioni dipende dal passo di discretizzazione scelto per

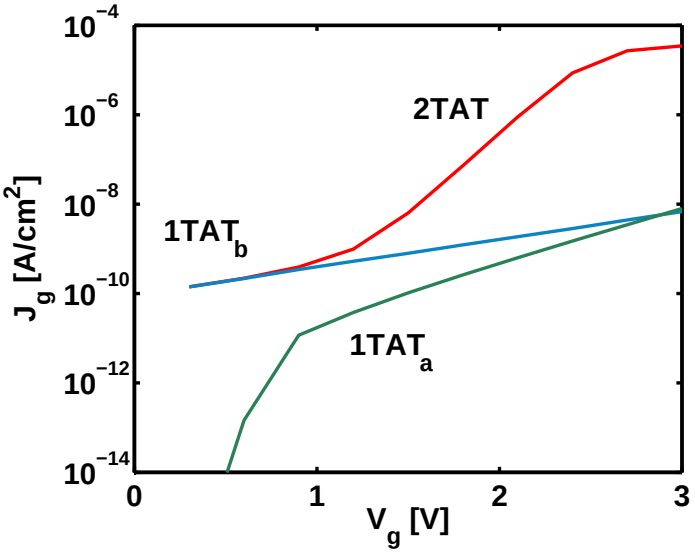


Figura 2.11: *Esempi di conduzione 1TAT e 2TAT. Le correnti 1TAT sono relative alle trappole che danno luogo al 2TAT prese singolarmente. L'interazione tra le trappole è evidente sopra 1V. I parametri che determinano queste correnti sono:* $t_{ox} = 3.7nm$, $x_{T,1} = 1.4nm$ *e* $x_{T,2} = 2.4nm$ *a partire dall'interfaccia* SiO_2*–bulk,* $E_{T,1} = 0eV$ *e* $E_{T,2} = 0.6eV$ *a partire dal fondo della banda di conduzione del bulk,* $\sigma = 10^{-16}cm^2$, $\tau_1 = 10^{-15}s$ *e* $N_{T,1} = N_{T,2} = 10^{10}cm^{-2}$.

il calcolo, questa volta necessariamente numerico e non analitico [31]. Per affrontare un problema del genere, è necessario prendere in considerazione dei metodi iterativi di tipo approssimato, come per esempio il metodo di Newton–Raphson, una generalizzazione del metodo di Newton in uno spazio funzionale multivariabile. Data la complessità computazionale di questo algoritmo, si è scelto di implementare unicamente il caso di densità deltiformi. Infatti, esse sono comunque in grado di ben approssimare i casi reali di conduzione 2TAT.

§2.3.4 Confronto con i dati sperimentali

Come banco di prova per il modello di conduzione 2TAT sin qui descritto, si sono messe a confronto le simulazioni numeriche con i dati sperimentali relativi ad un dispositivo MOS *High–k* provenienti da misure effettuate nell'ambito del progetto europeo NANOSIL, all'avanguardia nella ricerca nel campo della nanoelettronica. Il dispositivo in questione consiste dunque in un transistore n–MOS, con uno *stack* di dielettrici formato da circa $0.7nm$ di SiO_2 nativo e da $3nm$ di HfO_2, per uno spessore di ossido equivalente $EOT = 1.25nm$. Come contatto di *gate* si è utilizzato il TiN, mentre l'area è di $20\mu m \times 20\mu m$. Per

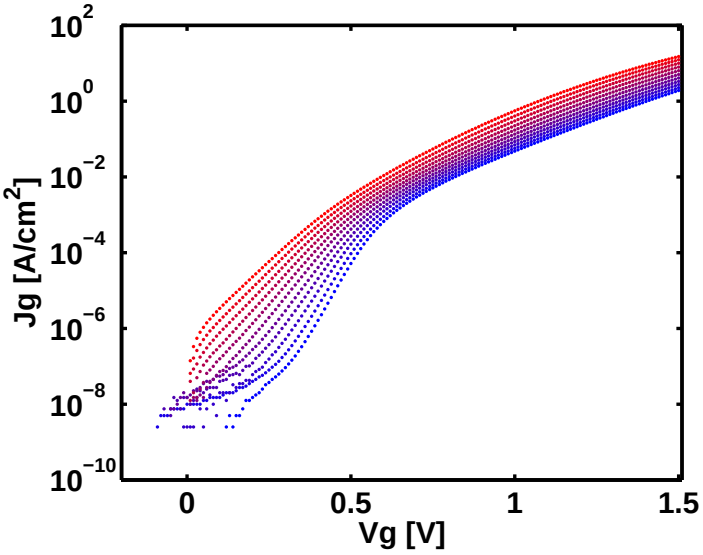


Figura 2.12: *Correnti di gate sperimentali del dispositivo campione. Sono mostrate le curve al variare della temperatura, da* $-50°C$ *(blu) a* $200°C$ *(rosso) con uno step di* $25°C$*.*

il nitruro di titanio si è utilizzato un valore della *work function* $\Phi_M = 4.5eV$, come riportato in [32], mentre per l'ossido di afnio si è utilizzato un valore della costante dielettrica $\epsilon_{HK} = 20$, del *band gap* $E_G = 5.8eV$, dell'*offset* della banda di conduzione nei confronti del silicio $\Delta E_{CB} = 1.8eV$ e della massa di *tunneling* efficace per gli elettroni $m_{HK} = 0.13m_0$, coerentemente a quanto riportato in [6][33]. Per quanto riguarda l'ossido si silicio, si è utilizzato $m_{ox} = 0.53m_0$. Di questo transistore è stata misurata la corrente di *gate* in un *range* di tensioni tra $0 \div 1.5V$, al variare della temperatura tra $-50°C$ e $200°C$, come mostrato in Fig. 2.12.

Per procedere al *fitting* dei dati sperimentali, si è usato come *software* di base il simulatore *QUASIMOD*, sviluppato all'interno del Diparimento di Elettronica e Informazione del Politecnico di Milano [34], che è in grado di fornire informazioni circa l'elettrostatica del dispositivo e le caratteristiche di quantizzazione nel canale. Agli algoritmi di calcolo della corrente di *tunnelig* diretto/Fowler–Nordheim e 1TAT, è stato aggiunto quello di calcolo della corrente 2TAT, necessario in quanto non è stato possibile giungere ad un *fitting* accettabile dei dati con i contributi di solo *tunneling* o 1TAT, soprattutto per quanto riguarda la variazione di corrente nel dominio della temperatura. Una buona interpolazione si è invece raggiunta considerando la corrente come una somma di contributi 2TAT, ciascuno con dei parametri spaziali ed energetici leggermente diversi, che vanno ad emulare così una densità di trappole continua funzione dello spazio e dell'energia $N_T \sim N_T(x_T, E_T)$. I valori di questi parametri sono riassunti nella Tab. 2.1. Le righe corrispondono alla coppia di

	$x_{1,T}[nm]$	$x_{2,T}[nm]$	$E_{1,T}[eV]$	$E_{2,T}[eV]$	$\sigma[cm^2]$	$N_{1,T}[cm^{-2}]$	$N_{2,T}[cm^{-2}]$
I	-0.55	-1.75	-0.1	0	10^{-16}	$2 \cdot 10^{10}$	$2 \cdot 10^{10}$
II	-0.55	-1.65	0	0.125	10^{-16}	$4 \cdot 10^{10}$	$4 \cdot 10^{10}$
III	-0.55	-1.65	0.1	0.3	10^{-16}	$1.8 \cdot 10^{11}$	$1.8 \cdot 10^{11}$
IV	-0.55	-1.65	0.2	0.5	10^{-16}	$1.4 \cdot 10^{12}$	$1.4 \cdot 10^{12}$
V	-0.55	-1.65	0.3	0.675	10^{-16}	$3 \cdot 10^{12}$	$3 \cdot 10^{12}$
VI	-0.55	-1.65	0.4	0.75	10^{-16}	$4 \cdot 10^{12}$	$4 \cdot 10^{12}$

Tabella 2.1: *Parametri di fitting per la corrente 2TAT nei confronti del dispositivo MOS–HK in esame.*

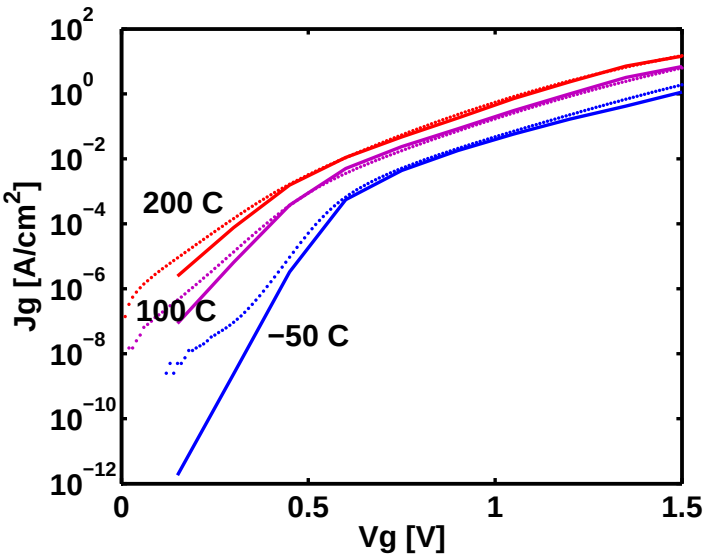


Figura 2.13: *Correnti di gate simulate (linee continue) e sperimentali (linee tratteggiate), per le temperature $-50°C$, $100°C$ e $200°C$.*

trappole prese in cosiderazione, $x_{1,T}$ e $x_{2,T}$ alla loro posizione spaziale a partire dall'interfaccia $SiO_2 - Si(bulk)$, $E_{1,T}$ e $E_{2,T}$ alla loro posizione energetica riferita all'energia del fondo della banda di conduzione del canale, *i.e.* al valore di tale banda all'interfaccia $SiO_2 - Si(bulk)$; σ è la *cross section* sia nel caso di interazione trappola–elettrodo che trappola–trappola mentre $N_{1,T}$ e $N_{2,T}$ sono le densità di difetti per unità di superficie, considerate quindi deltiformi in x ed E. Infine, si è scelto come valore di $\tau_{1,T} = \tau_{2,T} = 0.5 \cdot 10^{-15} s$

I risultati di queste simulazioni sono mostrati in Fig. 2.13, dove sono rappresentate le curve sperimentali e quelle simulate, per i valori di temperatura $-50°C$, $100°C$ e $200°C$. Come si può notare, il *fitting* è piuttosto accurato, e questo grazie alla precisa calibrazione dei parametri delle sei coppie di trappole prese in considerazione. In Fig. 2.14 è mostrata la caratteristica $I - V$ per la

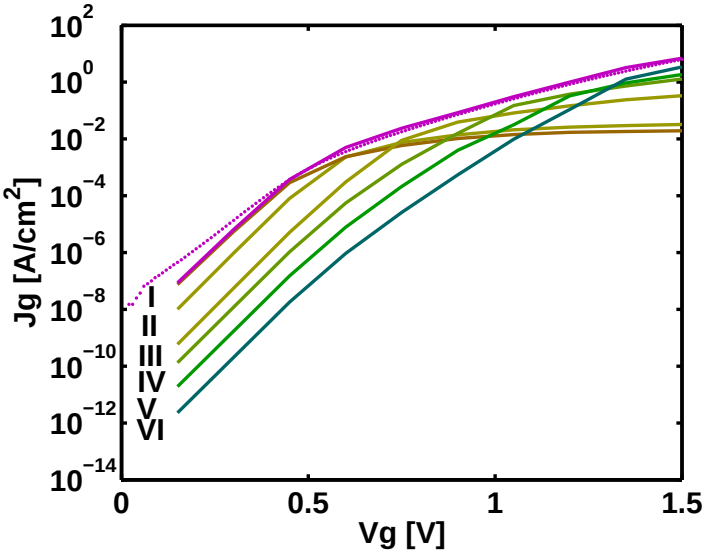


Figura 2.14: *Contributi delle diverse coppie di trappole ai fini della corrente totale. Il grafico fa riferimento al caso* $T = 100°C$.

sola temperatura $T = 100°C$, mettendo in luce come i contributi delle diverse coppie si sommino per creare la corrente totale. È interessante notare come ogni contributo sia dominante solo per un ristretto range di tensione: al variare del campo elettrico, cambiano le trappole che massimizzano il trasporto di carica.

Sembra dunque che si sia in presenza di un doppio strato di trappole: uno nell'ossido di silicio nativo, molto vicino all'interfaccia con l'*High–k*, e uno all'interno di quest'ultimo, ad una distanza di circa $1nm$ dall'interfaccia. Se spazialmente sono, in buona approssimazione, localizzate, così non è per le energie: a basse tensioni di *gate* dominano le trappole ad energia più bassa, mentre all'aumentare della tensione assumono maggior rilevanza le trappole ad energia più alta. Ciò è dovuto al fatto che più la barriera di potenziale si piega, *i.e.* la tensione sale, più la trasparenza di barriera aumenta per le trappole ad energia elevata, che "vedono" anche una maggior disponibilità di elettroni nel substrato dal momento che, rispetto al fondo della banda di conduzione del silicio di *bulk* all'interfaccia con l'ossido, le trappole si sono energeticamente abbassate, favorendo la cattura di elettroni.

Infine, consideriamo l'andamento della corrente totale al variare della temperatura. Se ipotiziamo un'andamento del tipo Arrhenius

$$J = J_0 \cdot e^{-\frac{E_a}{kT}}$$

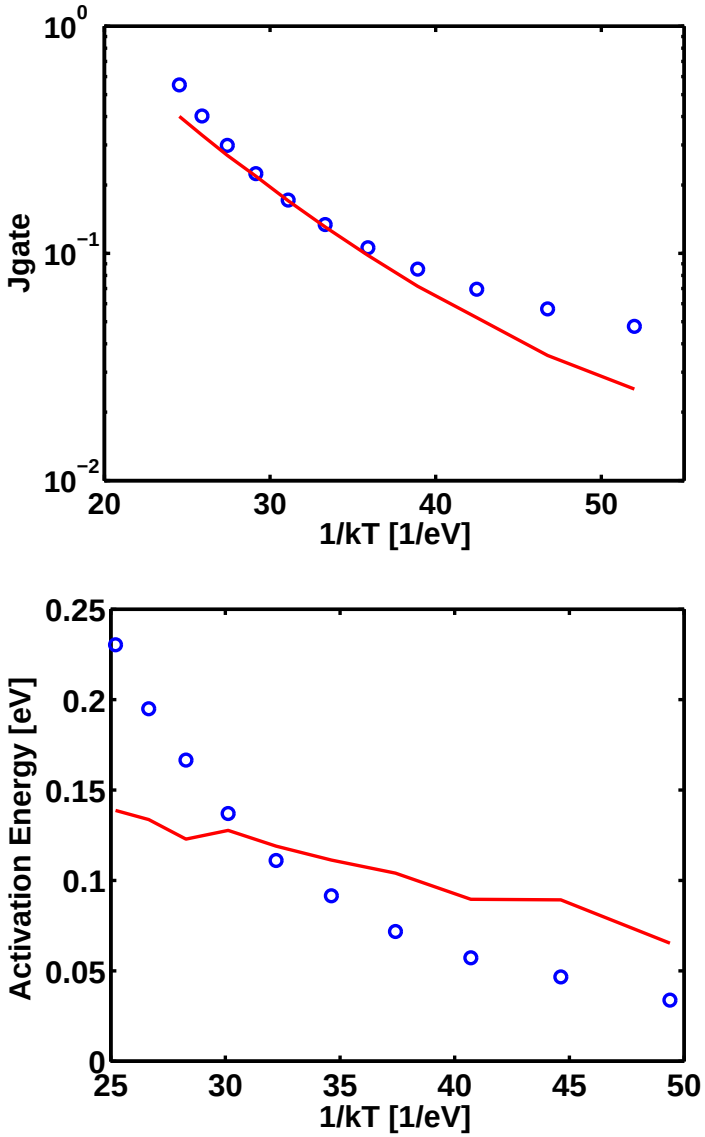


Figura 2.15: *Andamento della corrente di gate (a) e dell'energia di attivazione (b) in funzione di* $\frac{1}{kT}$*, alla tensione di gate di* $1V$*. Le linee tratteggiate rappresentano le misure sperimentali, quelle continue le curve simulate.*

con E_a che è detta *energia di attivazione*, alla tensione esempio di $1V$ la densità di corrente di *gate* e l'energia di attivazione hanno l'andamento rappresentato in Fig. 2.15, rispetto alla temperatura espressa in $\frac{1}{kT}$. Come si vede, l'andamento della densità di corrente simulata al variare della temperatura è sostanzialmente concordante con i relativi valori sperimentali, e l'energia di attivazione assume valori dell'ordine di $50meV \div 150meV$ nel caso simulato e di $50meV \div 250meV$ nel caso sperimentale.

§2.4 Dal monodimensionale al tridimensionale

Nelle precedenti sezioni abbiamo esposto il modello di conduzione TAT fino a 2 strati di trappole, in approssimazione monodimensionale. Questo ci obbligava a concepire i difetti sottoforma appunto di densità (nello spazio e nell'energia), perdendo così il carattere prettamente puntiforme che invece posseggono nella realtà. Il problema che insorge, con la miniaturizzazione dei dispositivi e l'accuratezza sempre maggiore dei processi produttivi, è che la realtà granulare dei fenomeni è sempre più manifesta, in maniera tale che non sempre una descrizione dei parametri da un punto di vista medio statistico, come lo è il caso della densità di trappole, è uno strumento ancora valido per la determinazione del corretto valore della corrente di *leakage* e di tutto ciò che da essa deriva. Si rende necessaria, pertanto, l'estensione del precedente modello alle tre dimensioni, almeno per quanto concerne la caratterizzazione dei difetti e la corrente ad essi imputabile. Non è stato possibile disporre di un simulatore in grado di provvedere l'elettrostatica e la quantizzazione per una struttura tridimensionale, che comprendesse quindi gli effetti di bordo e la possibile distorsione del potenziale dovuto alla eventuale carica elettrica dei difetti reticolari. Tuttavia si sono introdotte delle opportune modifiche negli algoritmi che tenessero in conto almeno la disposizione spaziale tridimensionale delle trappole, intese quindi d'ora in avanti puntiformi.

§2.4.1 Modello approssimato a N trappole

Come introdotto in precedenza, nel momento di passare da una situazione monodimensionale che descrive le trappole tramite le loro densità ad un ambiente tridimensionale che invece le descrive da un punto di vista "di singola trappola", è necessario trovare delle nuove espressioni per i flussi di carica che prescindano dal fattore densità di difetti.

Prendiamo in considerazione una situazione del tipo schematizzato in Fig. 2.16, in cui abbiamo rappresentato lo schema a bande di un dispositivo in presenza di una singola trappola localizzata in x_T e E_T. Ai fini del seguente ragionamento, le altre coordinate della trappola y_T e z_T sono, per il momento,

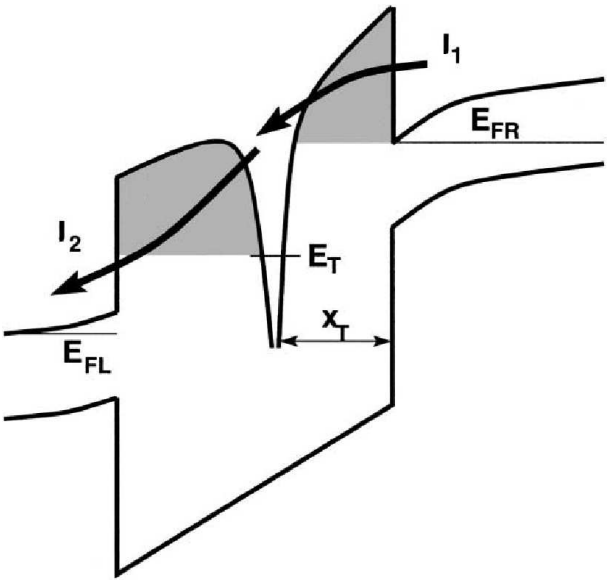


Figura 2.16: *Diagramma a bande nel caso 1TAT con indicati i rispettivi flussi di corrente [26].*

irrilevanti. Andiamo a considerare solamente i flussi di che verosimilmente sono dominanti ai fini della corrente e cioè quelli provenienti dal substrato e diretti verso il *gate*, in virtù del fatto che la tensione applicata a quest'ultimo è, nel nostro caso, positiva. Essi sono: il flusso tra substrato e trappola e il flusso tra la trappola e *gate* [35]

$$\begin{cases} I_1 = qc(1 - f_T) \\ I_2 = qe(f_T) \end{cases} \tag{2.55}$$

in cui I_1 e I_2 sono a tutti gli effetti delle correnti, c ed e sono i tassi di cattura e di emissione dal substrato e verso il *gate* (come calcolati nelle Eqs. 2.35, 2.36 e 2.40), rispettivamente e con f_T che rappresenta sempre la probabilità di occupazione della trappola. Notare come si tratti di una scrittura del tutto analoga a quella delle Eqs. 2.27 e 2.28, in cui sono scomparsi solamente i termini riferiti alle densità di trappole. Questo è consentito dal momento che, in precedenza, i prodotti $f_T \cdot N_T$ e $(1 - f_T) \cdot N_T$ rappresentavano il *numero* di difetti per unità di superficie occupati oppure liberi, e ora che sto considerando una singola trappola è sufficiente mantenere soltanto i termini relativi alla sua occupazione. In condizioni stazionarie, le due correnti si eguagliano, portando la f_T a valere

$$f_T = \frac{c}{c + e}$$

che se sostuita all'interno dell'Eq. 2.55 da come risultato

$$I_1 = I_2 = q\frac{c \cdot e}{c + e} \tag{2.56}$$

L'Eq. 2.56 afferma come la corrente che scorre tra *bulk* e *gate* dipenda dai tassi di cattura e di emissione, *i.e.* dai flussi di cattura e di emissione, in ragione di una loro media armonica. Infatti, il termine a destra dell'Eq. 2.56 può essere riscritto come

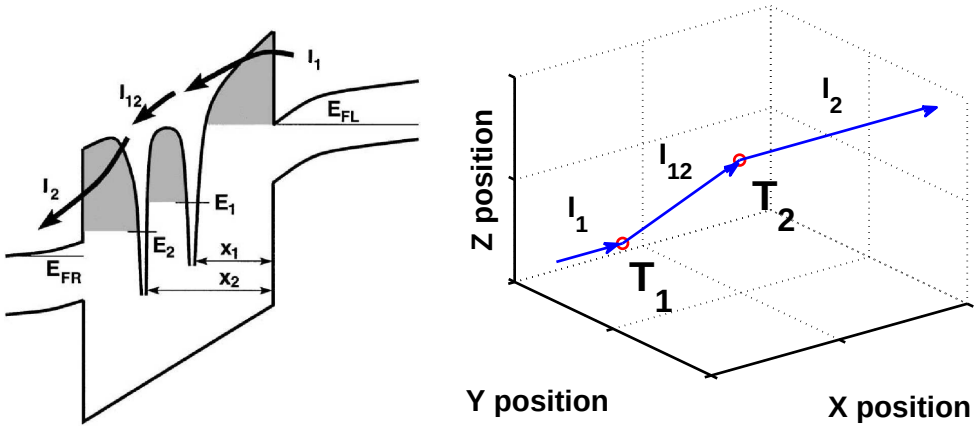
$$\frac{c \cdot e}{c + e} = \frac{1}{\frac{1}{c} + \frac{1}{e}}$$

Qualitativamente, questo risultato può essere spiegato concependo il flusso più debole come quello limitante il processo di conduzione [35]. Questo ragionamento può essere il punto di partenza per un'estensione del processo TAT ad un numero qualsivoglia di trappole, in cui il problema del calcolo rigoroso della funzione di occupazione diventa sempre più complicato fino a diventare inestricabile. Si avrebbe quindi un metodo, a questo punto approssimato, che non contempla più le proprietà di occupazione delle trappole e grazie al quale si può determinare, attraverso la semplice valutazione dei tassi di transizione tra trappola ed elettrodo o tra trappola e trappola, una stima ragionevolmente verosimile della corrente di perdita semplicemente attraverso una operazione di media armonica.

Un'altra cosiderazione va fatta in merito all'ipotesi iniziale dell'aver trascurato, in questo caso, i flussi di corrente emessi dalla trappola verso il substrato e di cattura da parte della trappola dal *gate*. Quest'ipotesi è tanto più valida quanto più polarizzato è il dipositivo, dal momento che è proprio la polarizzazione a sbilanciare, in maniera esponenziale, i corrispettivi flussi di cattura ed emissione. È sufficiente dunque prendere in considerazione solo il flusso di elettroni diretto dalla zona polarizzata negativamente a quella polarizzata positivamente.

Prendiamo ora in considerazione l'esempio di una conduzione 2TAT, rivista alla luce del precedente ragionamento. La situazione è schematizzata nella Fig. 2.17(a), in cui è rappresentato il diagramma a bande della struttura e le posizioni spaziali $x_{1,T}$ e $x_{2,T}$ ed energetiche $E_{1,T}$ e $E_{2,T}$ delle trappole, e nella Fig. 2.17(b) in cui sono mostrate anche le altre effettive (possibili) posizioni spaziali all'interno del dielettrico $y_{i,T}$ e $z_{i,T}$. Le espressioni approssimate per i flussi di corrente, già private cioè del contenuto informativo circa l'occupazione dello stato, sono

$$\begin{cases} I_1 = qc_{1,R} \\ I_{12} = qc_{21} \\ I_2 = qe_{2,L} \end{cases} \tag{2.57}$$



(a) (b)

Figura 2.17: *Diagramma a bande nel caso 2TAT con indicati i rispettivi flussi di corrente (a)[26] ed esempio di disposizione tridimensionale delle trappole nel dielettrico (b). $x = 0$ indica il substrato, $x = 1$ il gate.*

con il nuovo termine c_{21} che riassume le proprietà di interazione tra le trappole, come verrà spiegato più in dettaglio tra poco (vedi Sec. 2.4.2).

Stando a quanto detto per il caso 1TAT, quindi, la corrente totale 2TAT corrisponde alla media armonica dei singoli flussi approssimati

$$\frac{1}{I_{TOT}} = \frac{1}{I_1} + \frac{1}{I_{12}} + \frac{1}{I_2}$$

dove, ancora, si vede come sia il flusso più debole a risultare limitante ai fini della conduzione.

Prendendo, infine, in considerazione il caso più generale possibile, ovvero il caso in cui si sia in presenza di un percorso di TAT ad un numero di trappole N qualsiasi, il ragionamento è assolutamente analogo a quanto visto finora. I singoli flussi possono essere scritti come segue

$$\begin{cases} I_1 = qc_{1,R} & \text{substrato} \rightarrow \text{prima trappola}, \\ I_{i,i+1} = qc_{i,i+1} & (i+1)\text{–esima trappola} \rightarrow i\text{–esima trappola}, \; i \in [1, N-1], \\ I_N = qe_{N,L} & N\text{–esima trappola} \rightarrow \textit{gate} \end{cases}$$

mentre la corrente totale corrisponde ancora una volta alla media armonica dei singoli flussi

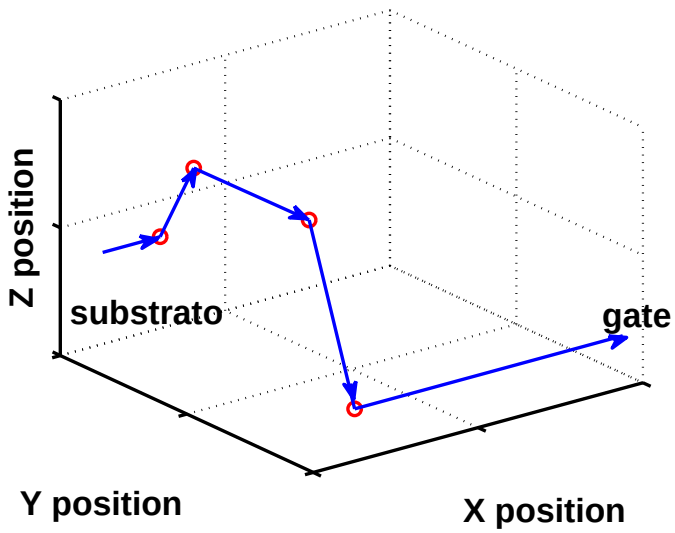


Figura 2.18: *Rappresentazione grafica di un percorso a quattro trappole tra substrato e gate*

$$\frac{1}{I_{TOT}} = \frac{1}{I_1} + \sum_{i=1}^{N-1} \frac{1}{I_{i,i+1}} + \frac{1}{I_N}$$

Nel caso di polarizzazioni negative, *i.e.* con un diagramma a bande invertito rispetto alla Fig. 2.17(a), è sufficiente andare a considerare i flussi di corrente nella direzione opposta, poiché saranno essi a dare il contributo dominante alla corrente totale. Vale la pena far notare, infine, come tralasciando i termini relativi alle funzioni occupazione, si consideri sempre "piena" la trappola da cui gli elettroni partono e sempre "vuota" quella in cui arrivano. Questa situazione approssimata si avvicina sempre più al vero man mano che la tensione applicata aumenta in modulo (sia essa positiva o negativa), tuttavia anche per tensioni non troppo distanti dallo zero si è riscontrato un buon accordo tra questo metodo e quello analiticamente esatto del capitolo precedente.

§2.4.2 Tasso di cattura inter–trappola

Se per quanto riguarda il calcolo dei tassi di cattura e di emissione tra elettrodo e una qualsiasi trappola non viene sostanzialmente modificata la teoria esposta nel capitolo precedente, così non è per quel che concerne i tassi di cattura inter–trappola. La loro natura puntiforme richiede una riscrittura dell'espressione in modo da poter eliminare i termini relativi alla densità di trappole (vedi l'Eq. 2.48), che non hanno più motivo di esistere dal momento

che sto considerando difetti singoli. Supponendo quindi di essere in presenza di 2 trappole, il tasso di cattura da parte della seconda di un portatore presente nella prima (utilizzato nell'Eq. 2.57), può essere riscritto come segue

$$c_{21} = \frac{\Omega}{4\pi}\frac{T_{12}}{\tau_1}$$

in cui Ω rappresenta l'angolo solido efficace di interazione tra i due difetti, rapportato a 4π in modo tale da fornire una stima della frazione dell'angolo sferico utile alla transizione. In particolare, Ω viene definito come

$$\Omega = \frac{\sigma_2}{r_{12}^2}$$

in cui σ_2 corrisponde, al solito, alla *cross section* della trappola di arrivo[3], mentre r_{12} identifica la distanza spaziale che intercorre tra i due difetti. Avendo entrambi le dimensioni di una superficie, è immediato individuare in Ω un angolo solido.

Non ultimo, anche il termine relativo alla trasparenza di barriera necessita di una sostanziale modifica, per tenere conto del percorso di transizione tridimensionale. Sempre utilizzando l'approssimazione semi–classica stile WKB, si può pensare di riscrivere la probabilità di attraversare la barriera in modo analogo al caso monodimensionale (vedi l'Eq. 2.11), a patto di considerare un diverso cammino spaziale

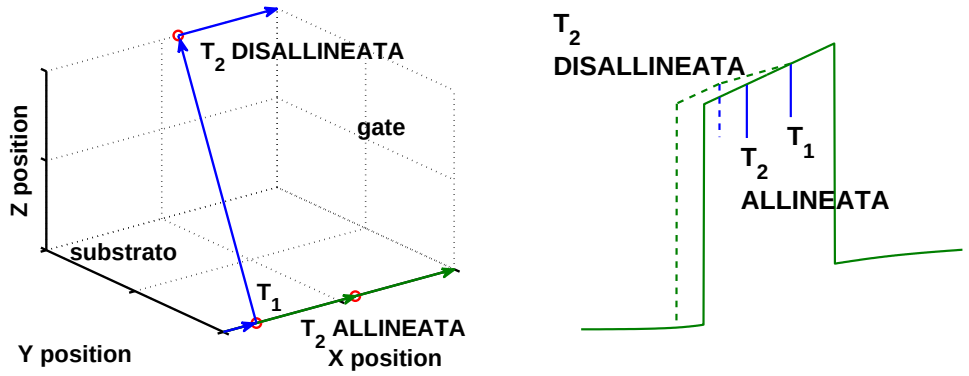
$$T_{12}(E) = \exp\left[-2\int_{\Gamma_0}^{\Gamma r_{12}} \frac{\sqrt{2\mu(U(\Gamma)-E)}}{\hbar}\mathrm{d}\Gamma\right]$$

in cui Γ rappresenta il segmento, nello spazio tridimensionale, congiungente le due trappole, mentre Γ_0 e $\Gamma_{r_{12}}$ sono i due estremi del segmento, *i.e.* le posizioni spaziali delle trappole stesse. A partire da $T_{12}(E)$ si può infine passare al termine T_{12} in maniera analoga al caso monodimensionale

$$T_{12} = \int_{E_i}^{+\infty} T_{12}(E)p(E)\mathrm{d}E$$

in cui E_i corrisponde sempre alla più elevata tra le energie delle trappole, al di sotto del quale non ho possibilità di transizione, e $p(E)$ che corrisponde alla funzione peso. Anche in questo caso possiamo scegliere una forma alla Maxwell–Boltzmann

[3]Vale la pena di far notare che si sta considerando il caso di transizione totalmente anelastica, nella quale la *cross section* assume un valore costante per tutti i valori di energia al di sopra di quello della trappola



(a) (b)

Figura 2.19: *In (a) è mostrata la possibile situazione in cui la seconda trappola è perfettamente allineata o meno alla prima, in (b) la barriera di potenziale equivalente nel primo caso (linea continua) e nel secondo (linea tratteggiata).*

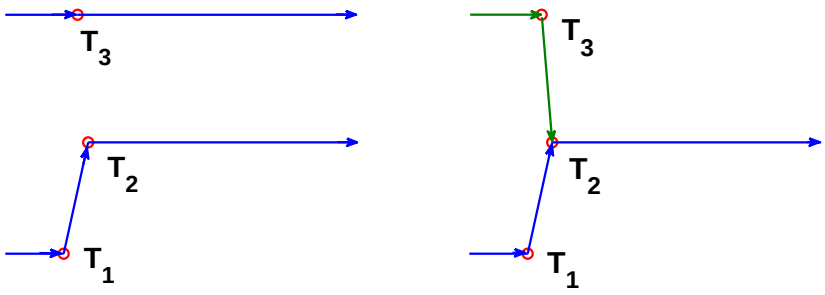
$$p(E) = e^{-\frac{(E-E_j)}{kT}} \frac{1}{kT}$$

dove il pedice j identifica la trappola di partenza.

È interessante constatare come un disallineamento della seconda trappola rispetto alla direzione di minor distanza della prima dall'elettrodo di arrivo provochi una diminuzione di entità esponenziale della trasparenza di barriera, dal momento che una transizione "obliqua" presenta uno spessore di barriera equivalente più cospicuo. La situazione è rappresentata schematicamente nelle Figs. 2.19(a) e 2.19(b). Questo effetto ha delle notevoli ripercussioni in merito alla scelta del percorso che un elettrone deve compiere nel momento di attraversare il dielettrico (come vedremo nella prossimo paragrafo): transizioni inter–trappola avverranno verosimilmente solo quando si sarà in presenza di un forte allineamento tra i difetti.

§2.4.3 Scelta del percorso TAT

Per completare la descrizione del modello tridimensionale in esame, è necessario approfondire il problema della scelta del percorso che un portatore verosimilmente segue nell'attraversare il dielettrico. È chiaro infatti che se siamo in presenza di N trappole, ma la loro distanza spaziale o energetica è tale che l'interazione reciproca è molto debole, esse di comporteranno sostanzialmente come N trappole singole, riducendo il fenomeno ad una semplice somma di 1TAT. Questo ovviamente è un caso estremo: in generale ci troveremo di



(a) (b)

Figura 2.20: *Rappresentazione schematica dei percorsi scelti dai due diversi algoritmi descritti. In (a) si nota come T_3 non "veda" T_2, che è stata eliminata assieme a T_1 poiché facenti parte del primo percorso, in (b) invece T_3 "vede" T_2, poiché solo T_1 è stata eliminata.*

fronte a percorsi costituiti da un numero di trappole variabili, *i.e.* diversi regimi di TAT, a seconda dei parametri e del numero complessivo delle trappole stesse. Bisogna quindi escogitare un algoritmo che permetta di individuare il percorso TAT più favorevole, cioè quello con maggior probabilità di essere seguito dai portatori. In linea di principio sarebbe necessario calcolare tutti i percorsi possibili e stilarne una graduatoria di importanza, andando poi ad eliminare, tra quelli meno efficienti, quelli con trappole in comune a quelli più alti in "classifica", ma questo comporterebbe uno sforzo computazionale non sempre accettabile, dal momento che il numero dei percorsi possibili cresce fattorialmente con il numero N di trappole. Pertanto si è dovuti ricorrere ad un metodo meno affidabile ma molto più semplice, che consiste in quanto segue. Per prima cosa, vengono calcolati i tassi di cattura da parte di tutte le trappole presenti rispetto all'elettrodo a più basso potenziale (*i.e.* da dove provengono elettroni). Determinato il maggiore tra essi, si considera la trappola che lo ha generato come la prima del percorso. Successivamente, si calcolano i flussi di corrente tra la trappola appena determinata e tutte le altre trappole, nonché con l'elettrodo a più alto potenziale (*i.e.* dove gli elettroni finiscono la loro corsa). Anche in questo caso vado a considerare il maggiore tra essi: se è relativo all'interazione con un'altra trappola, questa verrà considerata come la seconda del percorso ed il meccanismo di selezione si ripete, fissando quest'ultima come punto di partenza. Se invece invece il flusso più favorevole è quello corrispondente alla migrazione diretta verso l'altro elettrodo, allora il percorso può ritenersi concluso. A questo punto l'algoritmo si ripete, e qui si possono seguire due vie: la prima conduce semplicemente all'eliminazione di tutte le trappole coinvolte nel percorso di TAT appena calcolato, in modo

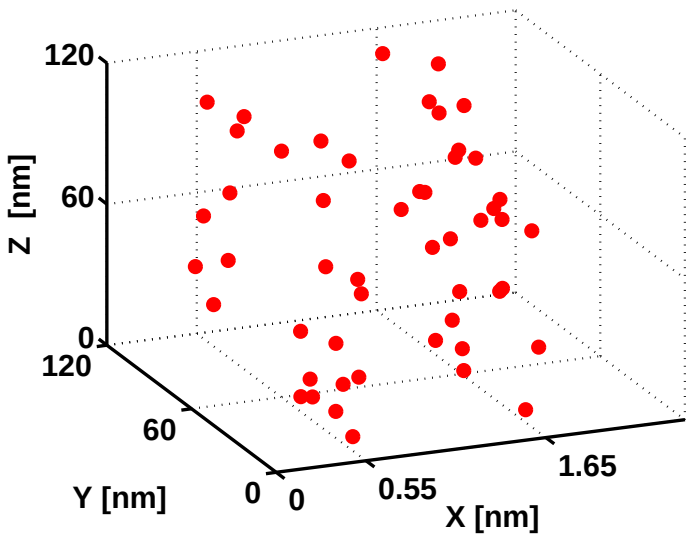


Figura 2.21: *Esempio di realizzazione di distribuzione spaziale del doppio strato di trappole per il dispositivo High–k.*

tale che il percorso successivo non includa nessuna delle trappole già utilizzate, la seconda invece conduce all'eliminazione soltanto della prima trappola del percorso appena calcolato, in modo tale da modificare solamente il primo *step* della conduzione e consentendo alle altre trappole di essere "riutilizzate". È evidente che nessuno dei due metodi può essere considerato, dal punto di vista teorico, superiore all'altro, e lo è altresì il fatto che entrambi possano portare a delle contraddizioni. Questo è dovuto alla natura approssimata del modello utilizzato, che rispecchia la nostra sostanziale ignoranza sulle funzioni di occupazione delle trappole, la valutazione delle quali, sì, garantirebbe un'esatta determinazione dei flussi di carica e quindi del percorso seguito con maggior probabilità. Nonostante ciò, i due algoritmi descritti sono da considerarsi sufficientemente accurati, a meno di situazioni patologiche puramente accademiche, e la discrepanza tra di essi è pure di lieve entità. Dovendo sceglierne uno, si è optato per il primo, dal momento che, eliminando un numero uguale o maggiore di trappole rispetto al secondo, risulta essere più veloce e meno impegnativo dal punto di vista delle risorse di calcolo. Nelle Figs. 2.20(a) e 2.20(b) sono rappresentati due esempi per gli algoritmi di calcolo descritti: nel primo si fa riferimento a quello adottato nel corso della presente trattazione, nel secondo a quello scartato.

§2.4.4 Validazione del modello *N*TAT

Come banco di prova per la validazione del modello tridimensionale *N*TAT è stato eseguito un confronto con i risultati delle simulazioni riportate nel Par.

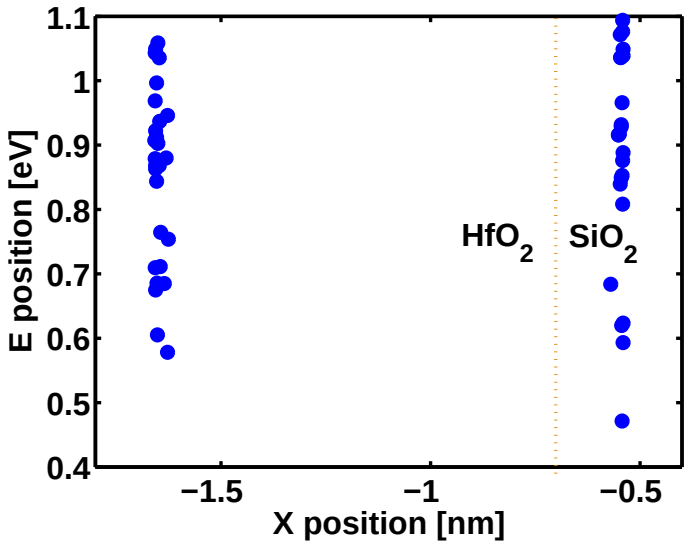


Figura 2.22: *Esempio di realizzazione di distribuzione nello spazio x_T ed energia del doppio strato di trappole per il dispositivo High–k. Lo zero per lo spazio corrisponde all'interfaccia SiO_2–bulk, per le energie al fondo della banda di conduzione del silicio alla medesima posizione.*

2.3.4, in cui si confrontavano a loro volta i risultati del modello monodimensionale 2TAT con i dati sperimentali relativi ad un dispositivo MOS *High–k*. Nel passare dall'una alle tre dimensioni si sono mantenuti gli stessi parametri fisici delle trappole (*cross section* e τ caratteristico, vedi Tab. 2.1), andando ad intervenire invece sul loro posizionamento spaziale ed energetico. Essendo infatti non più descritte in termini di densità per unità di superficie e di energia ma in termini puntuali, si è proceduto con dei meccanismi di generazione *random*[4] che riproducessero una distribuzione spaziale ed energetica analoga a quella monodimensionale in precedenza utilizzata. Trattandosi quindi di simulazioni con parametri estratti in modo casuale, è evidente che sono state necessarie un numero sufficiente di prove tali da consentire l'estrazione di un valor medio affidabile, non legato alla contingenza di una singola realizzazione. In Fig. 2.21 è mostrato un esempio di realizzazione per quanto riguarda la distribuzione spaziale delle trappole all'interno dei dielettrici: sono evidenti gli strati di difetti alle due diverse profondità in x_T, uno all'interno dell'ossido di silicio nativo e l'altro nell'ossido di afnio. In ciascuno dei due è evidente la natura stocastica della generazione delle trappole nelle dimensioni y_T e z_T, assunte, in assenza di motivi che lo controindicassero, distribuite uniformemente. In Fig.

[4]Per una comprensione più approfondita di tale argomenti si rimanda il lettore al prossimo capitolo.

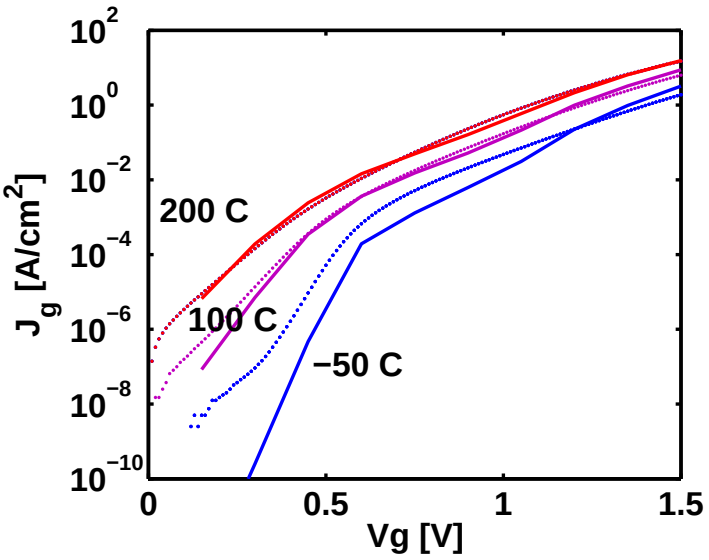


Figura 2.23: *Fitting delle curve sperimentali per il dispositivo High–k, con una media logaritmica su 10 realizzazioni.*

2.22 è mostrata la stessa realizzazione con, in ascissa, la posizione spaziale x_T e in ordinata l'energia della trappola, il cui valore è determinato anch'esso in maniera aleatoria a partire però da una distribuzione gaussiana, con media $\overline{E} = 1.1eV$ e deviazione standard $\sigma_E = 0.3eV$.

È di pregio il sostanziale accordo tra dati sperimentali e simulazioni tridimensionali secondo il modello NTAT. In Fig. 2.23 è mostrato il *fitting* delle curve sperimentali con le simulazioni effettuate. Le caratteristiche IV simulate derivano dalla media logaritmica delle caratteristiche di ciascuna realizzazione, anche se per ottenere questo buon accordo non è stato necessario svolgerne un numero elevato, anzi, una decina sono state sufficienti. Analizzando poi singolarmente ciascuna realizzazione, si può constatare come i percorsi 1TAT siano i più frequenti, anche se ai fini della corrente finale sono i percorsi 2TAT, che, seppur rari, dominano. Data l'ampiezza delle dimensioni trasversali del dispositivo rispetto allo spessore dei dielettrici, infatti, è abbastanza infrequente che due trappole siano sufficientemente allineate per dare vita al 2TAT, tuttavia, quando accade, la conduzione è molto efficiente e tende a sovrastare tutti gli altri contributi. La situazione è qualitativamente esposta in Fig. 2.24, in cui, per la medesima realizzazione citata in precedenza, sono mostrati alcuni percorsi sia di singola che di doppia trappola.

In conclusione, non rimane che constatare come il modello tridimensionale NTAT sia, malgrado le numerose approssimazioni che gli stanno alla base, sostanzialmente coerente col modello monodimensionale, in virtù della stessa capacità predittiva mostrata nei confronti del dispositivo preso in esame.

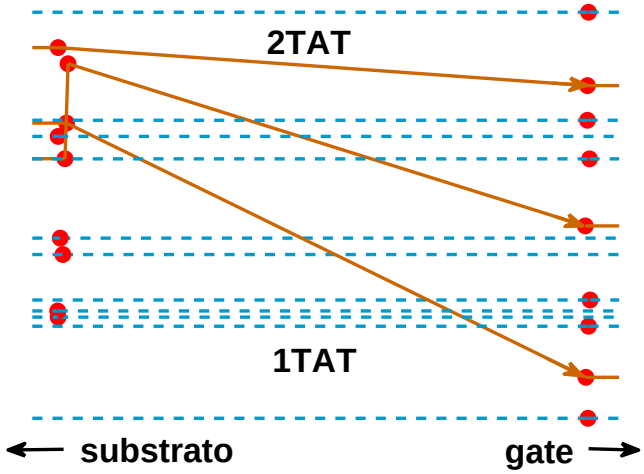


Figura 2.24: *Sezione bidimensionale (piano x–y) che mostra degli esempi di percorsi 1TAT (linee tratteggiate) 2TAT (linee continue) simulati nel dispositivo in esame. Gli ultimi danno i contributi dominanti ai fini della corrente totale, anche se più rari dei primi. A sinistra vi è il substrato, a destra l'elettrodo di gate.*

§2.5 Conclusioni

In questo capitolo sono stati presentati i modelli di conduzione per *tunneling* diretto/Fowler–Nordheim, per *Trap Assisted Tunneling* mediato da 1 e 2 strati di trappole in approssimazione monodimensionale, utilizzando per la trasparenza di barriera e i tassi di cattura/emissione la teoria WKB (approssimazione semiclassica), mentre per l'interazione con le trappole si è usato un modello statistico del tipo SHR. È stato poi mostrato un caso sperimentale di corrente di perdita imputabile al 2TAT e il corrispettivo *fitting* delle simulazioni. Successivamente è stato presentato il modello di conduzione NTAT tridimensionale, analizzandone le approssimazioni ed i limiti di validità. L'efficacia del modelli è stata verificata confrontando le sue predizioni con quelle del modello 2TAT rigoroso e con gli stessi dati sperimentali del caso precedente.

Capitolo **3**

Simulazioni Monte Carlo e statistiche della corrente di leakage

Any one who considers arithmetical methods of producing random digits is, of course, in a state of sin.

(Chiunque consideri metodi aritmetici per produrre cifre casuali è, naturalmente, in una condizione di peccato.)

John von Neumann

In questo capitolo si svolgerà un'analisi della statistica delle correnti di leakage *per i dispositivi MOS costituenti celle di memoria FLASH, focalizzandosi in particolare sulla dipendenza dallo spessore del* tunnel oxide*. I risultati saranno quindi messi a confronto con dei dati sperimentali. Le statistiche sono state estratte da simulazioni tridimensionali basate su metodi Monte Carlo specifici per eventi rari, utilizzando tecniche di generazione di numeri casuali di cui verrà data una rigorosa derivazione matematica.*

§3.1 Introduzione

Come abbiamo affermato nei Caps. 1 e 2, la miniaturizzazione dei dispositivi elettronici in generale e delle celle di memoria FLASH in particolare ha messo in luce il comportamento statistico di molte delle loro caratteristiche. Se nel caso di dispositivi di grandi dimensioni, quindi, gli effetti di questa aleatorietà in qualche modo si compensavano e ne consentivano uno studio basato su parametri medi, così più non è per i transistori ultrascalati, in cui le ridottissime dimensioni non sono sufficienti a consentire questi effetti di "compensazione". Bisogna quindi abbandonare una visione ristretta ad un singolo dispositivo per allargare l'orizzonte allo studio statistico su una collezione di essi, e valutare la distribuzione di probabilità di una certa grandezza che si vuole analizzare. Questa è la strategia adottata al momento di studiare il fenomeno chiamato *Stress Induced Leakage Current* (SILC), ovvero il problema della corrente di perdita attraverso il *tunnel oxide* causata dalla presenza di difetti all'interno di esso generati a loro volta dagli invasivi processi di programmazione/cancellazione della cella. Queste correnti TAT vengono, quindi, così analizzate: si prende in esame un campione sufficientemente rappresentativo di celle di memoria, come può esserlo un banco (detto *array*) e si procede alla determinazione della distribuzione di probabilità relativa alla corrente di *gate*. In questo modo si può stimare la probabilità che la corrente superi una certa soglia critica, determinata quest'ultima da considerazioni affidabilistiche, al variare del processo tecnologico e delle modalità di utilizzo del banco di memoria stesso. Quindi, se non possiamo sapere *quali* bit falliranno, possiamo almeno stimare *quanti* lo faranno. È chiaro che la coda di correnti oltre la soglia critica assume probabilità molto basse[1], sperabilmente inferiore a 10^{-9}, e questo crea dei problemi sia dal punto di vista sperimentale che numerico. Avendo dunque a disposizione una serie di dati sperimentali sulla statistica delle correnti di *gate* relative ad un *array* di celle [40], ci si è posti l'obiettivo di creare uno strumento di calcolo che fosse in grado di riprodurre tale distribuzione, al fine di determinare le proprietà, anch'esse di natura probabilistica, delle trappole che all'interno del *tunnel oxide* creano percorsi TAT, ai quali si imputa la corrente di perdita.

Si è pertanto proceduto alle simulazioni utilizzando tecniche cosiddette Monte Carlo, consistenti nel valutare la corrente di TAT per un numero K di dispositivi, attribuendo a ciascuno dei quali una diversa "realizzazione" per quando riguarda i parametri delle trappole e costruendo infine la distribuzione di probabilità risultante. Si coglia anche l'importanza dell'avere degli adeguati strumenti di generazione di variabili casuali: diverse distribuzioni di proba-

[1] I banchi di memoria attualmente in commercio superano il gigabit di capacità.

bilità per i parametri in ingresso impattano in maniera notevole sulla forma della statistica finale.

Durante il lavoro di simulazione, però, ci si è resi conto che con tecniche Monte Carlo standard si fatica ad ottenere una statistica affidabile per i valori di corrente più elevati. Ciò è dovuto al fatto che questa coda è imputabile ad un sottoinsieme delle celle con un numero di trappole anch'esso elevato, evento già molto raro di per sè dal momento che il numero medio di difetti per dispositivo è addirittura inferiore all'unità[2]. Tradotto in termini numerici, sono necessarie un numero elevatissimo di iterazioni per ottenere anche soltanto pochi campioni suscettibili di interesse, al punto di rendere la simulazione infattibile. Per rimediare a questo problema, peraltro molto noto in altri ambiti scientifici, si sono apportate delle modifiche all'impostazione stessa delle simulazioni, modifiche specifiche per eventi rari perché in qualche modo ne "amplificano" l'occorrenza [39].

Nella seguente sezione verrà innanzi tutto fornita una rigorosa derivazione matematica dei metodi di generazione di variabili stocastiche non–uniformi che sono state utilizzate per l'estrazione dei parametri aleatori degli stati trappola nel corso delle simulazioni Monte Carlo. Successivamente, verranno introdotte le principali tecniche di *enhancement* statistico, sempre da un punto divista strettamente formale. Infine, nell'ultima sezione, vedremo applicati questi strumenti matematici la caso di un *array* di celle FLASH.

§3.2 Sulla generazione di variabili casuali

Come si è gia osservato nella sezione precedente, l'importanza che riveste la generazione di variabili casuali nell'ambito delle simulazioni che adottano tecniche Monte Carlo è considerevole. Dalla loro accuratezza dipendono i risultati delle simulazioni stesse, dalla loro efficienza può dipendere la fattibilià stessa di una simulazione.

Dedichiamo pertanto questa sezione ad una breve rassegna dei metodi adottati nel corso del presente lavoro di tesi, metodi che costituiscono parte integrante delle implementazioni degli algoritmi utilizzati per la generazione dei parametri relativi alle trappole, quali posizione spaziale ed energetica, e che verranno sfruttati nelle simulazioni descritte più in dettaglio nelle prossime sezioni.

§3.2.1 Dall'uniforme al non uniforme

Molti dei *software* comunemente utilizzati nell'ambito della programmazione scientifica, tra i quali quelli utilizzati per l'implementazione degli algoritmi

[2]Ci si riferisce alla tecnologia e alle condizioni di utilizzo dell'*array* preso in esame.

in questo contesto di studio[3], sono in grado di fornire, tra le principali funzioni in libreria, un generatore di numeri *random* con distribuzione uniforme. Molteplici sono i metodi escogitati nel tempo per assicurare una distribuzione probabilistica il più possibile vicina all'uniformità, tuttavia occorre tener presente che, avendo a che fare con macchine la cui natura è algoritmica, qualsiasi speranza di ottenere delle variabili veramente casuali è destinata a spegnersi. Questo è il motivo per cui sarebbe più corretto riferirsi ad essi con il termine *pseudo–random*. Nonostante la loro natura prettamente deterministica, il grado di vicinanza ad una casualià genuina è spesso più che accettabile, ed è per questo che d'ora in avanti possiamo considerare come acquisita la generazione di numeri casuali secondo una distribuzione uniforme.

Il problema che si pone a questo punto è che spesso una distribuzione uniforme non è sufficiente a soddisfare le richieste di un certo modello matematico/fisico che si vuole simulare. Si rendono necessari quindi dei metodi che consentano di generare variabili secondo una prederminata distribuzione di probabilità, a partire dallo strumento di generazione *pseudo–random* uniforme. L'algoritmo più comune e di carattere più generale per soddisfare tale richiesta è, dal punto di vista teorico, molto semplice, e prende il nome di *Metodo dell' Inversione* o *Trasformata Integrale della Probabilità*. Esso presuppone solamente la conoscenza della funzione cumulativa della distribuzione che si desidera ottenere, detta F, e che sia continua e invertibile. Il metodo procede come segue [36]:

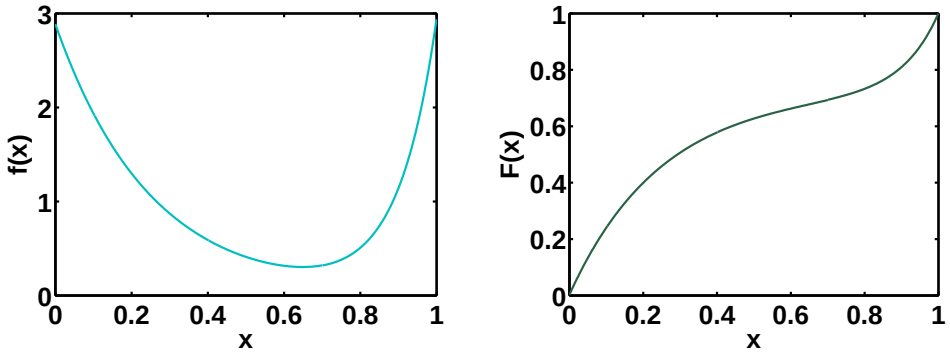
- Generare un numero casuale uniformemente distribuito, detto u;
- Calcolare il valore x tale che $F(x) = u$, chiamamo tale valore x';
- x' è il numero casuale distribuito secondo F.

La dimostrazione è anch'essa molto semplice e di seguito riportata

Dimostrazione. Assumiamo che F sia una distribuzione di ripartizione, continua, e che F^{-1} sia la sua inversa:

$$F^{-1}(u) = \inf \{x | F(x) = u, 0 < u < 1\}$$

[3]Vale la pena di citare il *Fortran* ed il *C*.



(a) (b)

Figura 3.1: *In (a) è mostrato l'andamento di $p(x)$ con parametri $a = 0$, $b = 1$, $\lambda_1 = 0.25$, $\lambda_2 = 0.1$ e $A = B = 2.88$. In (b) invece è mostrato l'andamento di $F(x)$.*

Andando a considerare quindi la distribuzione che assume x':

$$\begin{aligned} \Pr(F^{-1}(u) \leq x) &= \Pr(\inf\{x|F(x) = u\} \leq x) \\ &\quad \text{(per la definizione di } F^{-1}) \\ &= \Pr(u \leq F(x)) \\ &\quad \text{(applicando } F \text{ ad entrambi in membri)} \\ &= F(u) \\ &\quad \text{(dal momento che } u \text{ è uniforme)} \end{aligned}$$

□

Come si intuisce, il metodo è molto potente ed è il più generale possibile. Tuttavia non sempre è di semplice realizzazione pratica. La funzione cumulativa, infatti, potrebbe non essere analitica, ancor meno la sua inversa, constringendo il programmatore a soluzioni numeriche potenzialmente molto onerose dal punto di vista dello sforzo di calcolo e quindi, sostanzialmente, poco efficienti.

Per rendere l'idea, affrontiamo un caso pratico che si è presentato durante questo lavoro di tesi. L'obiettivo è dunque quello di generare dei numeri casuali che seguano una distribuzione di probabilità a "doppia" esponenziale

$$p(x) = Ae^{\frac{-x+a}{\lambda_1}} + Be^{\frac{x-b}{\lambda_2}} \tag{3.1}$$

la cui funzione cumulativa ha questa forma

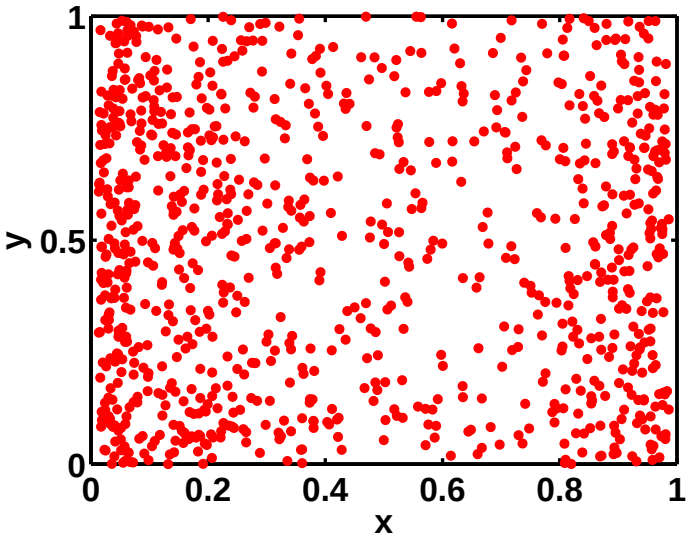


Figura 3.2: *Realizzazioni di una variabile casuale, x, che segue la distribuzione di probabilità rappresentata in Fig. 3.1(a). In y il campionamento è eseguito uniformemente.*

$$F(x) = A\lambda_1 \left(1 - e^{\frac{-x+a}{\lambda_1}}\right) + B\lambda_2 \left(e^{\frac{x-b}{\lambda_2}} - e^{\frac{a-b}{\lambda_2}}\right) \tag{3.2}$$

dove i prefattori A e B ottemperano alle esigenze di normalizzazione. Come si può constatare, $F(x)$ non è direttamente invertibile. Essendo pertanto infattibile l'operazione $F(u)^{-1} = x'$, si procede andando a cercare numericamente la radice dell'equazione

$$F(x') - u = 0$$

con uno dei tanti metodi esistenti per tale scopo. Tra i più comuni ricordiamo il metodo di bisezione, semplice e molto robusto, e quello di Newton, più veloce ma anche più delicato. In Fig. 3.1(a) è mostrata la funzione di probabilità $p(x)$ dell'Eq. 3.1, in Fig. 3.1(b) invece è mostrata la funzione cumulativa $F(x)$ relativa all'Eq. 3.2. In Fig. 3.2 è infine mostrata una possibile realizzazione: si può immediatamente constatare come, per la coordinata x, i punti si concentrino agli estremi del dominio, mentre siano più radi nella zona centrale, conformemente alla distribuzione di probabilità $p(x)$ a partire dalla quale sono stati generati. Per esigenze di visualizzazione il grafico è rappresentato in due dimensioni, in cui lungo y è stata scelta una distribuzione uniforme.

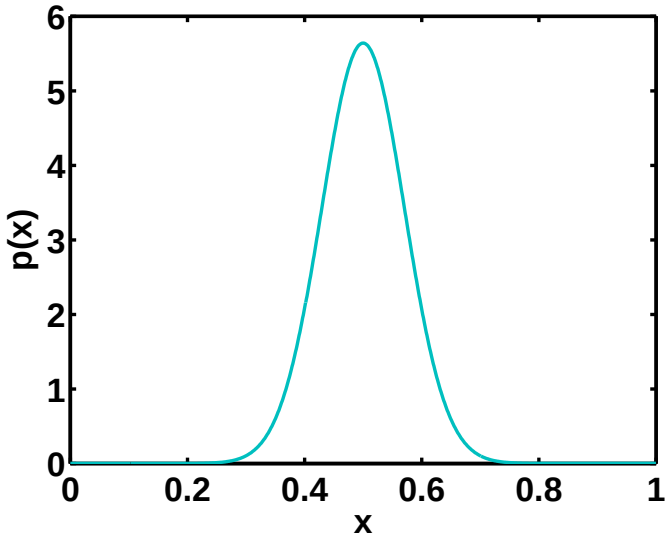


Figura 3.3: *Esempio di distribuzione gaussiana, con valor medio* 0.5 *e deviazione standard* 0.3.

§3.2.2 Distribuzione gaussiana

Ci sono particolari funzioni di probabilità che, grazie alla loro caratteristica forma funzionale, consentono di aggirare il metodo dell'inversione, lasciando spazio a metodi di generazione casuale *ad hoc* molto più efficienti, proprio perché specifici. Un esempio importantissimo di questo tipo di funzione è la distribuzione gaussiana, che ricorre spessissimo nei problemi di modellizzazione fisica di eventi probabilistici. Innanzi tutto, la distribuzione normale ha questa forma

$$p(x) = \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \tag{3.3}$$

in cui σ corrisponde alla deviazione standard e μ al punto medio. La sua funzione di ripartizione, non analitica, ha invece la forma

$$F(x) = \frac{1}{2}\left[1 + \operatorname{erf}\left(\frac{x-\mu}{\sqrt{2}\sigma}\right)\right]$$

Non essendo quindi possibile invertire questa funzione, per generare una variabile stocastica che segua una distribuzione normale si ricorre al seguente metodo, denominato *Trasformazione di Box–Muller* [37]:

Dimostrazione. Si considerino una coppia di variabili stocastiche (x, y), ciscuna

delle quali obbedisce ad una statistica gaussiana, e si assumano indipendenti tra loro. Pertanto, la densità di probabilità congiunta è

$$\begin{aligned} p(x,y) &= p(x)p(y) \\ &= \frac{1}{\sqrt{2\pi\sigma_x}} e^{-\frac{x^2}{2\sigma_x^2}} \cdot \frac{1}{\sqrt{2\pi\sigma_y}} e^{-\frac{y^2}{2\sigma_y^2}} \\ &= \frac{1}{\sqrt{2\pi\sigma_x\sigma_y}} e^{-\frac{x^2}{2\sigma_x^2} - \frac{y^2}{2\sigma_y^2}} \end{aligned}$$

Applicando una trasformazione di coordinate, per poter raccogliere all'esponente

$$\tilde{y} = y\frac{\sigma_x}{\sigma_y}, \qquad \mathrm{d}\tilde{y} = \mathrm{d}y\frac{\sigma_x}{\sigma_y}$$

in questo modo

$$p(x,\tilde{y}) = \frac{1}{\sqrt{2\pi\sigma_x\sigma_y}} e^{-\frac{x^2+\tilde{y}^2}{2\sigma_x^2}}$$

Dal momento che siamo in presenza di una simmetria radiale, viene naturale passare alle coordinate polari, anch'esse stocastiche, (R, Θ), definite come $0 \leq \Theta < 2\pi$, $x = R\cos(\Theta)$ e $\tilde{y} = R\sin(\Theta)$.
Chiaramente, Θ ha una distribuzione uniforme nell'intervallo $[0, 2\pi]$ e può essere campionata semplicemente come

$$\Theta = 2\pi \cdot u_1$$

Per quanto riguarda la componente radiale, invece, si scopre che si è in grado di integrare la densità di probabilità in modo analitico

$$\begin{aligned} \iint p(x,y)\mathrm{d}x\mathrm{d}y &= \iint p(x)p(\tilde{y})\mathrm{d}x\mathrm{d}\tilde{y}\frac{\sigma_x}{\sigma_y} \\ &= \iint \frac{1}{2\pi\sigma_x^2} e^{-\frac{x^2+\tilde{y}^2}{2\sigma_x^2}} \mathrm{d}x\mathrm{d}\tilde{y} \\ &= \iint \frac{1}{2\pi\sigma_x^2} e^{-\frac{r^2}{2\sigma_x^2}} r\mathrm{d}r\mathrm{d}\theta \end{aligned}$$

avendo fatto uso, nell'ultimo passaggio, dell'eguaglianza che lega i differenziali nelle due diverse geometrie $\mathrm{d}x\mathrm{d}\tilde{y} = r\mathrm{d}r\mathrm{d}\theta$.

Integrando la coordinata θ nell'intervallo $[0, 2\pi]$ e il raggio r tra $[0, R]$, ottengo come risultato

$$F(R, \Theta) = F(R) = 1 - e^{-\frac{R^2}{2\sigma_x^2}}$$

Applicando ora il metodo dell'inversione a questa funzione di ripartizione $F(R) = u_2$, si ottiene che

$$\begin{aligned} R &= \sqrt{-2\sigma_x^2 \ln(1 - u_2)} \\ &= \sqrt{-2\sigma_x^2 \ln(u_2)} \end{aligned}$$

in cui, nella seconda equazione, si è sfruttata la proprietà secondo cui se u_2 è distribuita uniformemente tra $[0, 1]$, lo è anche $1 - u_2$.

Se ora ritorniamo alle variabili trasformate, esse assumono il valore

$$\begin{aligned} x &= R\cos(\Theta) \\ \tilde{y} &= R\sin(\Theta) \end{aligned}$$

Esplicando il termine R e Θ e tornarno alla y originale

$$\begin{aligned} x &= \sqrt{-2\sigma_x^2 \ln(u_2)}\cos(2\pi u_1) \\ y &= \sqrt{-2\sigma_y^2 \ln(u_2)}\sin(2\pi u_1) \end{aligned} \tag{3.4}$$

□

Questa coppia di variabili (x, y) obbediscono entrambe ad una distribuzione normale, indipendentemente l'una dall'altra. Prendendo in considerazione quindi una sola delle due, si è in possesso della variabile stocastica richiesta all'origine.
In Fig. 3.3 è mostrato un esempio di distribuzione normale che vogliamo i nostri campioni numerici seguano, mentre in Fig. 3.4 sono mostrati tali campioni, Ancora una volta, per esigenze di chiarezza, la coordinata y è estratta uniformemente, mentre la distribuzione gaussiana è da ricercarsi nella coordinata x.

§3.2.3 Distribuzione poissoniana

Un'altra distribuzione ricorrente nei fenomeni fisici modellizzati in questo lavoro fa riferimento alla cosiddetta funzione di Poisson, intesa come limite di una Bernoulliana per un elevato numero di prove e una probabilità di successo esigua. Al contrario della gaussiana, che era una funzione continua in $\mathbb{R}$, la

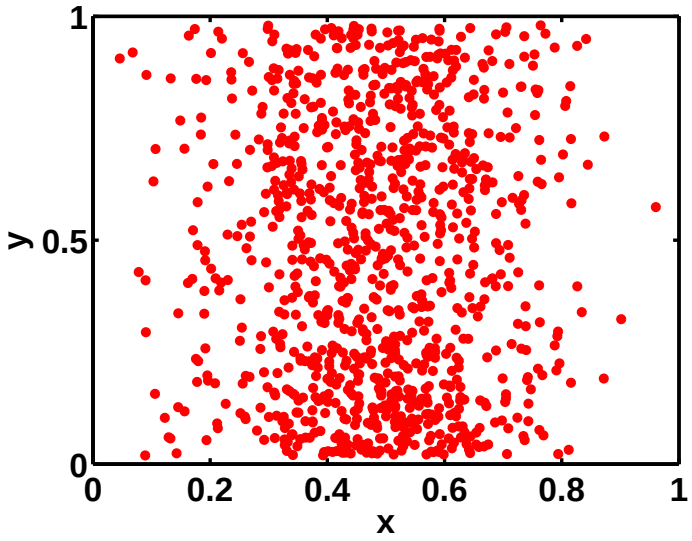


Figura 3.4: *Realizzazione di una variabile casuale che segue la distribuzione di probabilità normale rappresentata in Fig. 3.3*

poissoniana è invece di tipo discreto: la variabie stocastica n ad essa associata può assumere solamente valori naturali. La sua forma è

$$p(n) = \frac{e^{-\lambda}\lambda^n}{\lambda!}, \qquad \forall n \in \mathbb{N} \tag{3.5}$$

in cui λ rappresenta il valore medio, che nella poissoniana è anche uguale alla varianza.

Anche in questo caso la funzione di ripartizione non è direttamente invertibile, pertanto è necessario un metodo *ad hoc* che permetta di generare una variabile stocastica naturale che segua la distribuzione di Poisson. Un metodo molto ingegnoso, utilizzato poi nelle simulazioni descritte in questa tesi, è quello descritto da Donald Knuth [38]:

Condideriamo il valor medio λ come il prodotto di un tasso di conteggi nell'unità di tempo per un tempo di misura, ossia $\lambda = \lambda_0 \cdot T$.
La probabilità di non aver avuto ancora nessun conteggio, al tempo T, è pari a

$$\Pr(t_1 > T) = \Pr\left(n = 0|_T\right) = e^{-\lambda_0 T}$$

che, in funzione del tempo di musura T, assume la forma di un'esponenziale

decrescente[4]. Interpretando adesso questo tempo di misura alla stregua di una variabile casuale e volendo estrarne delle realizzazioni, è sufficiente utilizzare il metodo dell'inversione applicato alla funzione esponenziale, la cui funzione di ripartizione è $F(T) = 1 - e^{-\lambda_0 T}$. Dunque

$$T = \frac{1}{\lambda_0} \ln(u)$$

Aggiungiamo che il processo è senza memoria (*used is as good as new*), nel senso che una volta avvenuto il primo conteggio, il secondo avverrà indipendentemente da esso, come se fosse un "nuovo" primo conteggio.
Ora, immaginiamo di collezionare un numero di conteggi, ciascuno dei quali può essere descritto da una distribuzione di probabilità esponenziale decrescente, tale che la somma dei tempi necessari ad ottenerli sia minore del tempo di misura. Il numero di conteggi che ultimo soddisfa tale condizione risulta essere una realizzazione della variabile poissoniana che si stava cercando. In formule

$$\begin{gathered} \sum_{i=1}^{N} T_i < T \\ \sum_{i=1}^{N} -\tfrac{1}{\lambda_0} \ln(u_i) < T \\ \sum_{i=1}^{N} \ln(u_i) > -\lambda_0 T = \lambda \\ \prod_{i=1}^{N} u_i > e^{-\lambda} \end{gathered} \tag{3.6}$$

L'ultima equazione è quella definitiva, poiché mette in luce come un numero di natura poissoniana di valor medio λ possa essere generato a partire da una produttoria di numeri distribuiti uniformemente. In particolare, esso corrisponde all'ultimo numero tale per cui la condizione espressa nell'equazione risulta essere ancora verificata.
In Fig. 3.5 è mostrato qualche esempio di distribuzione poissoniana. Per valori elevati di λ essa si avvicina sempre più ad una distribuzione normale (come peraltro predicono i teoremi del limite centrale).

§3.2.4 Generazione correlata di trappole

Come sarà spiegato più in dettaglio nel prossimo capitolo, durante il lavoro di simulazione e del confronto con i risultati sperimentali ci si è resi conto che la posizione spaziale delle trappole risulta in qualche modo correlata alle posizioni delle altre eventuali trappole già esistenti. Ciò significa che, se per

[4]Per renderla una distribuzione di probabilità a tutti gli effetti occorre moltiplicarla per un termine di normalizzazione, in questo caso λ_0

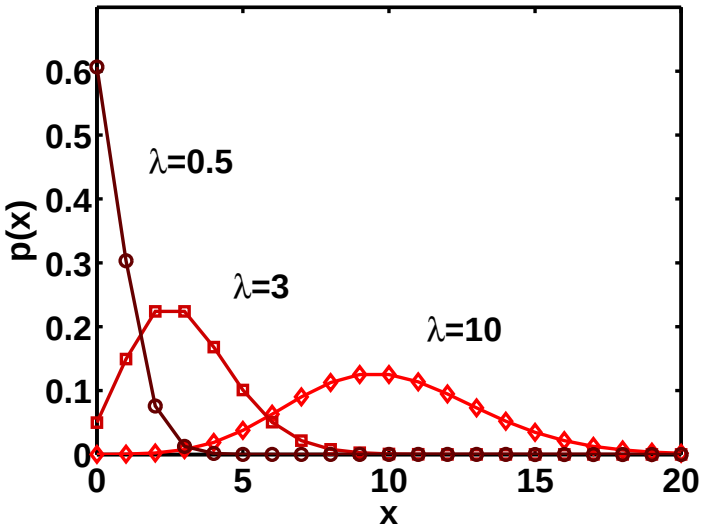


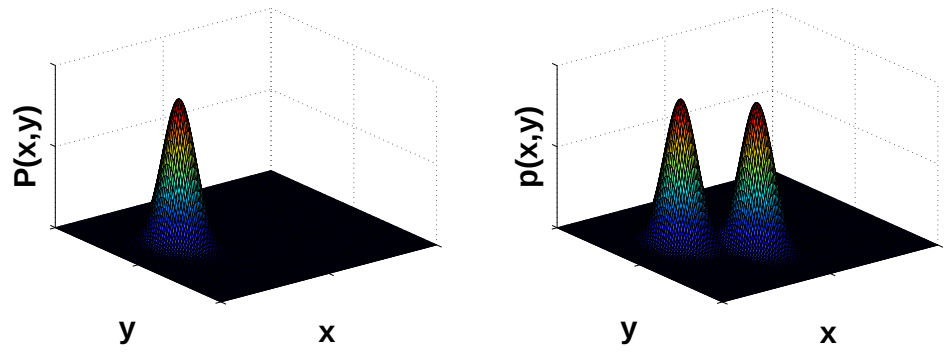
Figura 3.5: *Esempi di distribuzioni poissoniane con valor medio* $\lambda_1 = 0.5$, $\lambda_2 = 3$, $\lambda_1 = 10$.

esempio un primo difetto è stato generato in $\mathbf{r_1}$, la posizione del secondo sarà statisticamente legata al primo, $p(\mathbf{r_2}) = p(\mathbf{r_2})|_{\mathbf{r_1}}$, l'eventuale terzo sarà legato ad entrambe, $p(\mathbf{r_3}) = p(\mathbf{r_3})|_{\mathbf{r_1},\mathbf{r_2}}$, e così via per i successivi. È necessario allora escogitare un metodo che permetta di generare stocasticamente le posizioni delle trappole successive alla prima, fissato un criterio di correlazione. Assumiamo che ciascuna trappola modifichi la densità di probabilità spaziale semplicemente sommando a quella già esistente il suo contributo, in altre parole

$$p(\mathbf{r_3}) = p(\mathbf{r_3})|_{\mathbf{r_1},\mathbf{r_2}} = p(\mathbf{r_3})|_{\mathbf{r_1}} + p(\mathbf{r_3})|_{\mathbf{r_2}}$$

per il caso della terza trappola generata, e così per le successive. Nelle Fig. 3.6(a) e 3.6(b) sono mostrati degli esempi di densità di probabilità in un dominio (per semplicità) bidimensionale in presenza di una e due trappole, rispettivamente. Si è scelta una $p(\mathbf{r})|_{\mathbf{r}_i}$ di tipo gaussiano.

Il problema della generazione della nuova variabile casuale $\mathbf{r}$ sta nel fatto che $p(\mathbf{r})$ è per l'appunto una funzione delle 3 coordinate spaziali, e in generale non presenta alcuna simmetria significativa. Non potendo contare su strumenti creati *ad hoc*, proprio a causa della generalità della funzione, è necessario ricorrere al metodo dell'inversione. Per poterlo fare è però necessario ridurre il numero di variabili indipendenti da tre ad una, dato che il teorema 3.2.1 è assicurato solo in quest'ultimo caso. Si può pensare quindi di parametrizzare le tre dimensioni con una quarta, che diventa a questo punto l'unica indipendente



(a) (b)

Figura 3.6: *esempi di densità di probabilità in un dominio bidimensionale in presenza di una (a) e due (b) trappole.*

$$\mathbf{r} = (x, y, z) \longrightarrow \mathbf{r(t)} = (\mathbf{x(t)}, \mathbf{y(t)}, \mathbf{z(t)})$$

Se a questo punto facciamo corrispondere la parametrizzazione delle tre variabili ad un percorso $\Gamma(t)$ nello spazio tale che questo sia in grado di ricoprire tutto il dominio, la funzione di ripartizione passa da questa forma

$$F(X, Y, Z) = \Pr(x \leq X, y \leq Y, z \leq Z)$$

a quest'altra

$$F(T) = \Pr(t \leq T) = \int_0^{\Gamma(T)} p(\mathbf{r}(T'))\mathrm{d}\Gamma(T')$$

che è un semplice integrale di linea. In Fig. 3.7 è mostrata la funzione di ripartizione della probabilità in Fig.3.6(b) in funzione della variabile parametrica T, assumendo come percorso Γ una linea tale che, in un dominio bidimensionale, percorre l'asse delle ascisse a fissata ordinata, per poi aggiornare il valore di quest'ultima e ripetere l'operazione.
È facile ora procedere con il metodo dell'inversione, ponendo $F(T) - u = 0$ e trovando la sua radice T_0 per via numerica. A questo punto è sufficiente andare a valutare le tre coodinate spaziali nel punto T_0 e si ottiene una terna di valori coerenti con la distribuzione di probabilità $p(\mathbf{r})$

$$\begin{cases} x_0 = x(T_0) \\ y_0 = y(T_0) \\ z_0 = z(T_0) \end{cases}$$

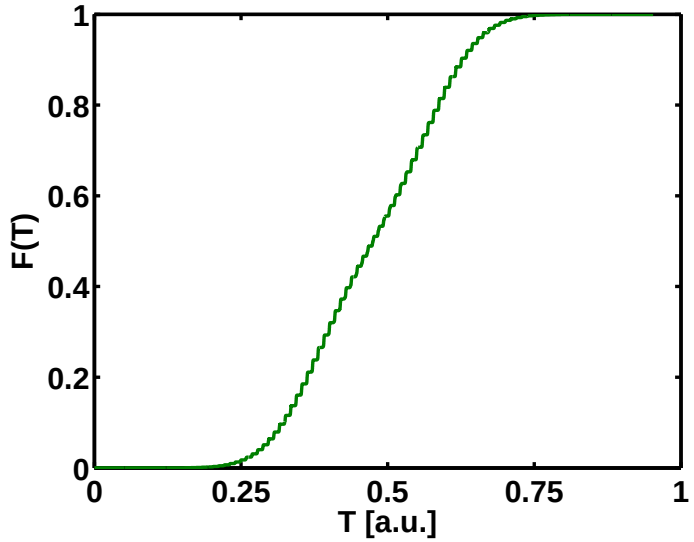


Figura 3.7: *Funzione di ripartizione della probabilità in Fig.3.6(b) in funzione della variabile parametrica T, assumendo come percorso Γ una linea tale che, in un dominio bidimensionale, percorre l'asse delle ascisse a fissata ordinata, per poi incrementare il valore di quest'ultima e ripetere l'operazione.*

In in Fig. 3.8 sono rappresentate alcune realizzazioni, ottenute con tale metodo, per la densità di probabilità bidimensionale della Fig. 3.6(b). Ognuna di esse potrebbe rappresentare la posizione di un'eventuale terza trappola.

Una cosiderazione finale. Sebbene molto potente, questo metodo è molto dispendioso in termini di potenza di calcolo, poiché deve lavorare su una griglia tridimensionale. Pertanto, in molte delle simulazioni presentate nei prossimi capitoli si è fatto uso, qualora fosse stato necessario generare trappole correlate, di un metodo approssimato che consiste nel considerare il baricentro delle trappole esistenti come il centro di una funzione densità di probabilità a simmetria sferica. In questo modo è stato possibile ricorrere ad uno dei metodi descritti neli precedenti paragrafi, che sono invece molto efficienti. Il disaccordo tra il metodo approssimato e quello rigoroso si è rivelato inconsistente nel momento in cui il numero di trappole non superi la decina, condizione peraltro rispettata in tutte le simulazioni effettuate.

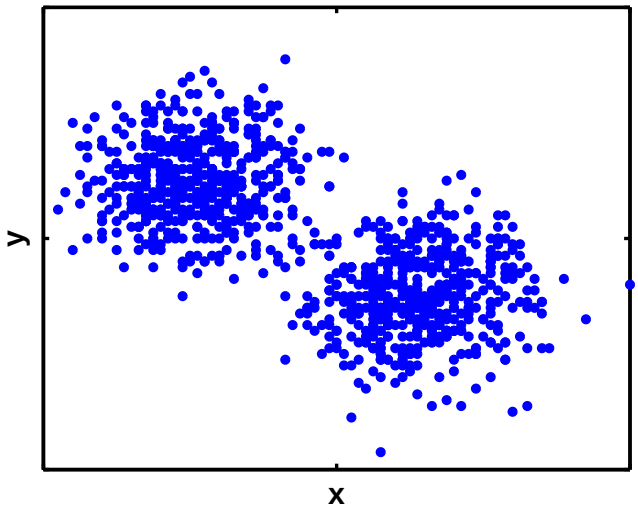


Figura 3.8: *Esempi di campionamento per la densità di probabilità bidimensionale della Fig. 3.6(b). Ognuno dei punti rappresenta rappresenta la possibile posizione di un'eventuale terza trappola.*

§3.3 Tecniche Monte Carlo per Eventi Rari

§3.3.1 Concetti di base

Innanzi tutto, introduciamo i concetti base relativi alle simulazioni Monte Carlo standard. Supponiamo di voler conoscere la probabilità γ che avvenga un certo evento A. Un modello del sistema è simulato n volte, in modo tale da essere in possesso di n campioni, e per ogni realizzazione si registra se l'evento A è accaduto oppure no. Normalmente, i campioni sono tra loro stocasticamente indipendenti. Quindi, se consideriamo la variabile casuale X, che sarà di tipo Bernoulliano dal momento che $X_i = 1$ se l'evento A accade all'i–esima iterazione, $X_i = 0$ viceversa, una stima di γ è

$$\hat{\gamma} = \sum_{i=1}^{n} \frac{X_i}{n}$$

Consideriamo inoltre che $E(X_i) = \gamma$ per definizione di γ e che $Var(X_i) = \gamma(1-\gamma) = \sigma^2$ poiché Bernoulliana. Applicando il teorema del limite centrale, che dice che $\sum_{i=1}^{n} X_i$ è approssimativamente una distribuzione normale, se n è sufficientemente grande, il valore dello *stimatore* $\hat{\gamma}$ si avvicina al suo vero valore in ragione di quanto segue.

Poniamo $E(\sum_{i=1}^{n} X_i) = n\gamma$ e $Var(\sum_{i=1}^{n} X_i) = n\sigma^2$, pertanto la variabile casuale

$$Z = \frac{\sum_{i=1}^{n} X_i - n\gamma}{\sqrt{n\sigma^2}}$$

ha media 0 e varianza 1. Il teorema del limite centrale afferma quindi che se $n \to +\infty$ allora $Z \to N(0,1)$, con $N(0,1)$ intesa come la distribuzione normale a media 0 e varianza 1. Ricordando che $\Pr(-z \leq Z \leq z) = \mathrm{erf}(z)$, si ottiene che

$$\Pr\left(\gamma \in \left(\hat{\gamma} - \frac{z\sigma}{\sqrt{n}}, \hat{\gamma} + \frac{z\sigma}{\sqrt{n}}\right)\right) = \mathrm{erf}(z)$$

in cui l'intervallo $I = \left(\hat{\gamma} \mp \frac{z\sigma}{\sqrt{n}}\right)$ è chiamato *intervallo di confidenza* per γ. Nei casi pratici, poiché stimiamo γ con $\hat{\gamma}$ ma σ^2 è sconosciuta, stimiamo quest'ultima come $\hat{\sigma}^2 = n\hat{\gamma}(1-\hat{\gamma})/n$.

La velocità di convergenza è quindi misurata in basa all'ampiezza dell'intervallo di confidenza, che vale $2z\sigma n^{-1/2}$. Come si vede esso decresce in ragione dell'inverso della radice quadrata del numero di campioni. Ora, supponiamo che A sia un evento raro, o equivalentemente $\gamma \ll 1$. Per numeri sì piccoli, l'errore assoluto dato dalla misura dell'intervallo di confidenza non è di sufficiente interesse: l'accuratezza del processo di simulazione è invece valutata attraverso l'*errore relativo* $RE = z\sigma n^{-1/2}/\gamma$. Questo ci porta direttamente al problema principale nell'ambito degli eventi rari, dal momento che

$$RE = z\frac{\sqrt{\gamma(1-\gamma)}}{\sqrt{n}\gamma} \approx \frac{z}{\sqrt{n}\sqrt{\gamma}}$$

Ovvero che se vogliamo garantire un certo errore relativo, per eventi la cui probabilità tende a zero, dobbiamo incrementare il numero di campioni

$$n = \frac{z}{RE^2\gamma}$$

Ovviamente, non sempre è possibile aumentare a piacimento il numero di simulazioni: i tempi di calcolo si dilatano in maniera ingestibile, tempi per la maggior parte sprecati nell'elaborazione di situazioni il cui interesse è irrilevante.

Esistono quindi varie tecniche che consentono di aggirare i limiti imposti da queste considerazioni. Le due famiglie più importanti prendono il nome di *Importance Sampling* (IS) e *Splitting Tecnique*.

§3.3.2 Importance Sampling

L'*Importance Sampling* è probabilmente l'approccio più famoso per l'analisi degli eventi rari. Esso riduce la varianza dello stimatore, incrementando anche l'occorrenza dei suddetti eventi. L'idea generale alla base del metodo è quella di cambiare la legge di probabilità del sistema in studio, al fine di campionare più frequentemente quegli eventi che sono più "importanti" per la statistica finale. Ovviamente, i risultati grezzi della simulazione devono essere riportati nei termini della misura iniziale. Questo è ottenuto mediante l'utilizzo di una funzione chiamata *likelihood ratio*, rapporto di verosimiglianza.

L'impostazione generale è come segue. Per esigenze di maggior generalità rispetto al precedente paragrafo, assumiamo che il sistema sia rappresentato da una certa variabile casuale X, e che l'obiettivo sia il valore di aspettazione γ di una certa funzione Ψ di X : $\gamma = E(\Psi(X))$, dove $\gamma \ll 1$. L'indicatore A si riferisce all'evento preso in considerazione. Assumiamo anche che X sia una variabile casuale di tipo reale, la cui densità è indicata con f, che Ψ rapprensenti l'evento di tipo bernoullinano ($\Psi = 0$ l'evento non si verifica, $\Psi = 1$ l'evento si verifica) dipendente da X e che la varianza di $\Psi(X)$ sia σ^2, finita. Lo stimatore standard di γ è

$$\hat{\gamma} = \frac{\sum_{i=1}^{n} \Psi(X_i)}{n}$$

dove le X_i corrispondono a dei campioni indipendenti di X secondo f.
L'*Importance Sampling* consiste nel campionare X da una diversa densità di probabilità $\tilde{f}$, chiamando questa trasformazione *cambio di misura*, con la sola condizione che $\tilde{f} > 0$ se $\Psi(x)f(x) > 0$. Aggiungiamo a pedice del valore di aspettazione l'indicazione sul tipo di densità che si sta utilizzando, E_f o $E_{\tilde{f}}$. Ovviamente, in generale, $\gamma \neq E_{\tilde{f}}(\Psi(X))$. Infatti

$$\gamma = \int \Psi(x)f(x)\mathrm{d}x = \int \Psi(x)\frac{f(x)}{\tilde{f}(x)}\tilde{f}(x)\mathrm{d}x \tag{3.7}$$

che possiamo anche riscrivere come $\gamma = E_{\tilde{f}}(\Psi(x)L(x))$ dove L è definita come $L(x) = f(x)/\tilde{f}(x)$ ed è chiamato rapporto di verosimiglianza. Se campioniamo n copie di X utilizzando la densità $\tilde{f}$ e mediamo i valori ottenuti considerando il rapporto di verosimiglianza, otteniamo un nuovo stimatore di γ, che chiamiamo $\tilde{\gamma}$

$$\tilde{\gamma} = \frac{\sum_{i=1}^{n} \Psi(X_i)L(X_i)}{n}$$

mentre la varianza assume la forma

$$\tilde{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} \Psi^2(X_i)L^2(X_i) - \tilde{\gamma}^2$$

Come si può intuire, una buona densità $\tilde{f}$ è tale per cui $\tilde{f} \gg f$ per opportuni valori di x, ad esempio $x \in A$.

Confrontiamo ora l'ampiezza degli intervalli di confidenza del metodo Monte Carlo standard e IS, mettendo in luce come la varianza dello stimatore sia, dopo il cambio di misura, inferiore al caso originale. A questo scopo, riferiamoci direttamente alle rispettive varianze esatte (assumendo che i loro estimatori siano sufficientemente precisi)

$$\begin{aligned}
Var(\tilde{\gamma}) &= \frac{1}{n} Var_{\tilde{f}}\left(\Psi(X)L(X)\right) \\
&= \frac{1}{n}\left[E_{\tilde{f}}\left(\Psi^2(X)L^2(X)\right) - \gamma^2\right] \\
&\ll \frac{1}{n}\left[E_{\tilde{f}}\left(\Psi^2(X)L(X)\right) - \gamma^2\right] \\
&= \frac{1}{n}\left[\int \Psi^2(x)\frac{f(x)}{\tilde{f}(x)}\tilde{f}(x)\mathrm{d}x - \gamma^2\right] \qquad (3.8) \\
&= \frac{1}{n}\left[\int \Psi^2(x)f(x)\mathrm{d}x - \gamma^2\right] \qquad (3.9) \\
&= \frac{1}{n}\left[E_f(\Psi^2(X)) - \gamma^2\right] \\
&= \frac{1}{n} Var_f(\Psi(X)) \\
&= Var(\hat{\gamma})
\end{aligned}$$

Questa è l'idea base del metodo IS. Scegliendo un buon cambio di misura si hanno quindi dei miglioramenti significativi in termini di efficienza, derivanti dal fatto che lo stimatore $\tilde{\gamma}$ è molto più accurato di $\hat{\gamma}$. Non è sempre banale, tuttavia, individuare una densità $\tilde{f}$ che consenta questa soluzioni. Esiste la soluzione ottima, se consideriamo una funzione costruita in tal modo $\tilde{f}(x) = f(x)\Psi(x)/\gamma$, con il rapporto di verosimiglianza che diventa $L(x) = \gamma/\Psi(x)$ e dunque

$$Var(\tilde{\gamma}) = \frac{1}{n} Var_{\tilde{f}}(L(X)\Psi(X)) = \frac{1}{n} Var_{\tilde{f}}[\gamma] = 0$$

anche se, come si può vedere, presuppone la conoscenza del valore di γ, che è invece il nostro obiettivo. Tuttavia questo ragionamento ci rassicura sul fatto

che esistano dei cambi di misura molto potenti, il che rende il metodo IS tra i più utilizzati in quest'ambito. Vedremo più avanti in questo capitolo come usarlo in modo astuto nel caso delle simulazioni delle correnti TAT.

Un'ultima considerazione in merito: nel caso in cui la variabile casuale X fosse discreta e non continua, tutto l'impianto continua ovviamente a valere. Formalmente, nelle Eqs. 3.7, 3.8 e 3.9 si dovrà sostituire il termine integrale con il termine sommatoria per rendere conto del cambio di dominio.

§3.3.3 Splitting Tecnique

La *Slitting Tecnique* è basata su un'idea completamente diversa rispetto all'IS. Ora, infatti, non cambiamo la legge di probabilità che guida il modello, ma utiliziamo un meccanismo di selezione delle "traiettorie" che sembrano portare ai rari eventi. L'idea principale è quella di decomporre i percorsi verso gli eventi di interesse in sottopercorsi la cui probabilità non sia troppo troppo esigua, incoraggiando le realizzazioni che imboccano questi sottopercorsi mediante una replicazione, scoraggiando invece quelle che seguono percorsi di non interesse semplicemente terminandole. I sottopercorsi sono generalmente delimitati da "livelli". Partendo da un livello prestabilito, le realizzazioni del processo che non raggiungono il livello successivo verso l'evento raro vengono "uccise", quelle che invece lo fanno vengono "clonate", da cui il termine *splitting*, e ogni copia prosegue indipendentemente dalle altre. Questo crea una deriva artificiale verso l'evento di interesse favorendo le traiettorie che vanno nella giusta direzione. Alla fine, lo stimatore obiettivo della simulazione può essere ricostruito semplicemente moltiplicando il contributo di ogni traiettoria per un opportuno fattore di peso.

Le maggiori difficoltà di questa tecnica stanno evidentemente nel definire i livelli che determinano i sottopercorsi, il loro numero e anche il fattore di moltiplicazione/estinzione per le traiettorie che superano o non superano i determinati livelli.

In termini matematici, nel modo più generale possibile, la *Splitting Tecnique* può essere vista come segue. Supponiamo che un certo processo stocastico X rappresenti il sistema sotto esame, e che esso sia incluso nello spazio degli stati S. L'obiettivo della simulazione sta nel trovare il valore di aspettazione γ tale che $\gamma = \Pr(\tau_A < \tau_0)$, dove $\tau_0 = \inf\{t > 0 : X(t) = 0, X(t^- \neq 0)\}$ e $\tau_A = \inf\{t > 0 : X(t) \in A\}$ per A che è un sottoinsieme di S raramente visitato da X. Consideriamo una sequenza di K sottoinsiemi di stati inclusi $A_1 \supset A_2 \supset \ldots \supset A_K = A$, con lo stato iniziale $0 \notin A_1$. Ora, partendo la simulazione da 0, se e quando la traiettoria raggiunge un qualche stato $s_1 \subset A_1$, allora sono costruite n_1 copie di X, tutte nello stato s_1 e che evolvono in maniera indipendente. Lo stesso procedimento è applicato quando qualsiasi delle versioni di X create raggiunge uno stato $s_2 \in A_2$, clonandone ciascuna in n_2 copie, e così via per A_3, ecc, fino ad arrivare (sperabilmente) a $A = A_K$. Per

questa ragione, gli A_K sono chiamati "livelli". Una volta arrivati in A, il metodo viene considerato concluso. Sostanzialmente, la dinamica stocastica di X è mantenuta invariata (non ho effettuato cambi di misura), ma facendo diverse copie delle traiettorie "buone" incrementiamo le probabilità di raggiungere l'evento raro A. Ovviamente, bisogna decidere cosa fare delle traiettorie che non hanno raggiunto A, ma questo dipende da quale sottocategoria di metodo di *Splitting* si sta utilizzando.

Per ottenere lo stimatore, siano $\tau_{A_i} = \inf\{t > 0 : X(t) \in A_i\}$ e $p_i = \Pr(\tau_{A_i} < \tau_0 | \tau_{A_{i-1}} < \tau_0)$, con $i = 1, 2, \ldots, K$ e $p_1 = \Pr(\tau_{A_1} < \tau_0)$. Pertanto $\gamma = \prod_{i=1}^{K} p_K$. Al fine di ottenere un numero significativo di traiettorie che intersecano A, è necessario che le probabilità p_i non siano troppo piccole. Esplicitando lo stimatore, essendo H il numero di eventi che raggiungono A

$$\tilde{\gamma} = \frac{H}{\prod_{i=0}^{K-1}}$$

Per quanto riguarda la varianza, invece, dopo un po' di algebra e qualche semplificazione

$$Var(\tilde{\gamma}) = \gamma^2 \sum_{i=1}^{K} \frac{1 - p_i}{\prod_{j=1}^{i} p_j n_{j-1}}$$

L'errore relativo è, come visto prima, proporzionale alla radice quadrata della varianza. Da questa espressione si potrebbe quindi trarre la conclusione che sia sufficiente aumentare il numero di livelli K e il numero di cloni n_i per migliorare la stima sul valore di aspettazione, ma si tratta di una osservazione parziale: con il loro aumento anche il tempo di calcolo cresce. Per questo è d'obbligo un'accurata e non banale analisi di tutti i fattori di costo nel momento di procedere con questa tecnica. Nel prossimo paragrafo sarà spiegato come si sia tentato di utilizzarla nelle simulazioni delle correnti di TAT, e del perchè ad essa sia stata preferita la tecnica dell'*Importance Sampling*.

§3.4 Risultati numerici e confronto con i dati sperimentali

Le misure sperimentali in nostro possesso provengono dall'analisi di un *array* di memorie FLASH della dimensione di $512kb$, dall'area di dispositivo di $0.06\mu m^2$ e dagli spessori del *tunnel oxide* $t_{ox} = 6.5$, 8.8 e $9.7nm$. Tutto l'*array* è stato sottoposto a un numero di cicli di programmazione/cancellazione pari a 10^4, attraverso una programmazione *Channel Hot Electrons* e una cancellazione Fowler–Nordheim uniforme. L'estrapolazione della corrente di perdita avviene mediante la misura periodica della tensione di soglia V_T di ogni singola cella, applicata una tensione positiva al *gate*, e valutandone l'evoluzione

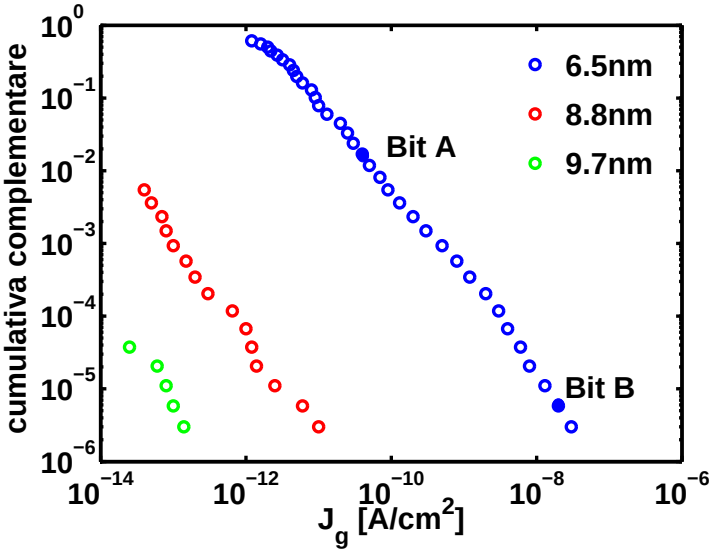


Figura 3.9: *Distribuzioni cumulative complementari per i tre diversi spessori di tunnel oxide [40]. Bit A e Bit B rappresentano le celle selezionate per l'array con* $t_{ox} = 6.5nm$.

temporale. Misurando lo *shift* della tensione di soglia in un tempo noto, si è in grado di calcolare la quantità di carica trasferita e quindi la corrente di *leakage*. Per le misure, si è scelta una tensione di *floating gate* $V_{FG} = 4V$ [40].

In Fig. 3.9 sono mostrate le distribuzioni cumulative complementari, ovvero 1−funzione cumulativa (detta anche funzione di sopravvivenza) per i tre diversi spessori di ossido. Come si può notare, per t_{ox} decrescenti la dispersione statistica aumenta sensibilmente. Inoltre, per il caso $t_{ox} = 6.5nm$, la parte iniziale della curva mostra una diversa pendenza dovuta all'instaurarsi del *tunneling* FN come meccanismo di conduzione dominante.

Prima di procedere ad una simulazione che rendesse conto della statistica complessiva, si è scelto di indagare in maniera più approfondita alcune singole celle, che in Fig. 3.9 abbiamo chiamato Bit A e Bit B, per individuare il regime di conduzione nelle diverse zone della curva. Infatti il solo meccanismo 1TAT non è sufficiente a spiegare uno *spread* delle correnti così ampio: bisogna ricorrere almeno al 2TAT, ovvero bisogna presupporre che i difetti all'interno dell'ossido siano, in qualche modo, cooperanti. In Fig. 3.10 sono mostrate le caratteristiche IV sperimentali e quelle simulate per le suddette celle. Da notare come il comportamento del Bit A sia giustificato da un meccanismo 1TAT, mentre quello del Bit B da uno di tipo 2TAT. Il valore dei parametri interpolanti è di seguito riportato: massa efficace elettronica nel diossido di silicio $m_{ox} = 0.53m_0$, *cross section* delle trappole $\sigma = 4 \cdot 10^{-10} cm^2$, tempo caratteristico delle trappole $\tau_{att} = 10^{-16} s$. La posizione spaziale della trappola relativa al Bit A è $x_T = 4.32nm$ dall'interfaccia ossido–semiconduttore, mentre

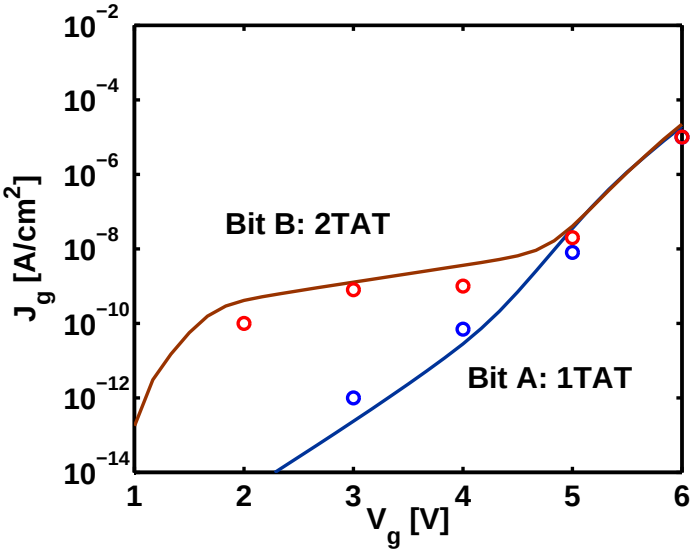


Figura 3.10: *Caratteristiche IV sperimentali (pallini) e simulate (lienee continue) per le celle Bit A e Bit B. Esse sono generate da un tipo di conduzione* $1TAT$ *e* $2TAT$*, rispettivamente.*

quella energetica è $E_T = -0.5eV$ rispetto al fondo della banda di conduzione del substrato all'interfaccia. Relativamente la Bit B, invece, $x_{T,1} = 2.9nm$ e $x_{T,2} = 5.6nm$ mentre $E_{T,1} = -0.7eV$ e $E_{T,2} = -0.3eV$. Da notare che il valore estremamente elevato per quanto concerne la *cross section* è da ricercarsi nella carica elettrica positiva posseduta dallo stato trappola, che distorcendo il potenziale attorno ad essa offre una minor barriera di potenziale per gli elettroni, diminuzione sintetizzata appunto nell'incremento della σ, da intendersi come *efficace*.

Attestato che siamo in presenza di difetti cooperanti, procediamo quindi con l'esporre i risultati relativi alla simulazione Monte Carlo principale. Innanzi tutto, il modello di conduzione utilizzato è quello tridimensionale NTAT, con i parametri relativi alle trappole generati in questo modo:

- *Cross section* e tempo caratteristico delle trappole deterministico. Essi valgono $\sigma = 4 \cdot 10^{-10} cm^2$ e $\tau_{att} = 10^{-15} s$.

- Numero di trappole per cella che segue una statistica poissoniana (vedi Par. 3.2.3), di valor medio $N = N_T \cdot Vol_{ox}$, dove con N_T si è indicata la densità media di trappole per unità di volume e con Vol_x il volume del *tunnel oxide*. N_T è un parametro fissato e vale $10^{15} cm^{-3}$, mentre Vol_x è un parametro variabile.

- Distribuzione spaziale uniforme, sia lungo la direzione longitudinale x che sulle trasversali y e z.
- Distribuzione energetica di tipo gaussiano (vedi Par. 3.2.2), con valor medio $\bar{E} = -1.2eV$ e deviazione standard $\sigma_E = 0.5eV$.

Queste distribuzioni spaziali ed energetiche adottate sono quelle che, a conti fatti, garantiscono il *best fitting* dei dati sperimentali.

Purtroppo, se dovessimo limitarci all'utilizzo di una simulazione Monte Carlo standard, è chiaro che non si riuscirebbe a scendere sotto probabilità dell'ordine di 10^{-5}, dal momento che i tempi di calcolo, al di sopra delle 10^5 iterazioni, iniziano a diventare proibitivi. Saremmo in possesso dunque di uno strumento numerico che non è nemmeno in grado di raggiungere la precisione dei dati sperimentali, men che meno di andare oltre. Sarebbe infatti opportuno riuscire a scendere ulteriormente fino a probabilità dell'ordine di 10^{-9}, dal momento che i banchi di memoria sono attualmente in grado di raggiungere delle capacità di gigabits, e il fallimento anche di un solo bit porta alla compromissione di tutto l'*array*. È chiaro quindi che bisogna introdurre qualche sistema che sia in grado di valutare più accuratamente gli eventi rari, che sono quelli responsabili delle alte correnti di nostro interesse, a discapito degli eventi più comuni dei quali invece ci è concesso avere una conoscenza meno approfondita. In nostro soccorso, si è intuito, vengono i metodi descritti nella sezione precedente, resta solo da individuare il corretto modo di utilizzo. Per individuarlo, si è fatto il seguente ragionamento.
Dal momento che il numero medio di trappole per singola cella risulta piutosto basso (0.4 per $tox = 6.5nm$ e $N_t = 10^{15}cm^{-3}$ [40]), è chiaro la maggior parte delle realizzazioni presenteranno nessun o un difetto[5], in seguito alla poissonianità del processo. Ma come abbiamo visto in precedenza, la coda di alte correnti è data sostanzialmente da celle in cui sono presenti almeno due difetti, per lo più cooperanti. L'idea a questo punto è quella di diminuire i casi in cui siano presenti nessuna o una trappola, che sprecano risorse di calcolo inutilmente, per aumentare invece artificialmente l'occorrenza di casi in cui il numero delle trappole è superiore. Questo corrisponde al meccanismo di *Importance Sampling* vero e proprio: si modifica la statistica dei parametri in ingresso in modo da mettere in luce alcuni particolari eventi (rari), in questo caso la presenza di un numero di difetti maggiore di uno. Ovviamente, ai fini della statistica finale, occorre tenere in conto del cambio di misura effettuato, attraverso il rapporto di verosimiglianza. In pratica, si opera *forzando* il numero di trappole ad assumere un particolare valore, per poter effettuare una serie di realizzazioni con quelle precise condizioni e tanto accurate quanto si vuole (nei limiti della macchina, s'intende).

[5] $\Pr(N \geq 2) \sim 6\%$, secondo la statistica di Poisson.

Più formalmente, la situazione si può modellizzare nel seguente modo. Poniamoci innanzi tutto in un ambito Monte Carlo standard. Coerentemente al Par. 3.3.2, sia X_i il numero di difetti per celle alla i–esima realizzazione e sia $\Psi(X_i)$ l'evento "l'i–esima corrente di perdita è compresa in un certo intervallo" (per la costruzione della distribuzione cumulativa). Lo stimatore risultante, per ogni intervallo di corrente, è semplicemente dato dalla somma delle $\Psi(X_i)$, che ricordiamo assume valore 0 o 1, diviso il numero delle realizzazioni, poniamo n. Ciò equivale ad affermare che l'esito di ogni realizzazione viene pesata per $1/n$, percui il valore di probabilità più piccolo raggiungibile è direttamente dipendente dal numero di prove effettuate: più l'evento è raro più prove sono necessarie, e, se è troppo raro, l'evento risulterà addirittura "invisibile". In un conteso di *Importance Sampling*, invece, si può agire come segue. Si effettua una sola prova con $X = 0$, ossia $n_0 = 1$, dato che in questo caso la corrente di perdita è data dal *tunneling* FN, per noi deterministico. Si procedo poi con un numero n_1 di prove con $X = 1$ e con un numero n_2 di prove con $X \geq 2$ (in cui la statistica interna è sempre poissoniana, solamente le $X < 2$ sono scartate a priori). Sia $n_{TOT} = n_0 + n_1 + n_2$ il numero totale di simulazioni effettuate. In questo modo, il rapporto di verosimiglianza da associare alle varie realizzazioni diventa:

$$\begin{cases} \Psi(0) \longrightarrow L(0) = \frac{f(0)}{\tilde{f}(0)} = poiss(0)\frac{n_{TOT}}{n_0} \\ \\ \Psi(1) \longrightarrow L(1) = \frac{f(1)}{\tilde{f}(1)} = poiss(1)\frac{n_{TOT}}{n_1} \\ \\ \Psi(2) \longrightarrow L(2) = \frac{f(X\geq 2)}{\tilde{f}(X\geq 2)} = poiss(X \geq 2)\frac{n_{TOT}}{n_2} \end{cases}$$

ed il peso al momento del conteggio, invece (è sufficiente dividere per il numero totale di simulazioni)

$$\begin{cases} w(0) = poiss(0) \\ \\ w(1) = \frac{poiss(1)}{n_1} \\ \\ w(2) = \frac{poiss(X\geq 2)}{n_2} \end{cases}$$

da cui si evince come la più piccola probabilità raggiungibile non dipenda più dal numero totale di realizzazioni, ma, nel nostro caso, da n_2. Ovviamente sarà un valore di probabilità estremamente inferiore al caso standard, dato che il termine $poiss(X \geq 2)$ al numeratore abbatte sensibilmente il peso tanto più quanto l'evento è raro. Effettuando dunque un cambio di misura iniziale, modificando cioè la statistica dei parametri in ingresso, si è riusciti a slegare il numero totale di simulazioni dal peso associato al momento del conteggio a

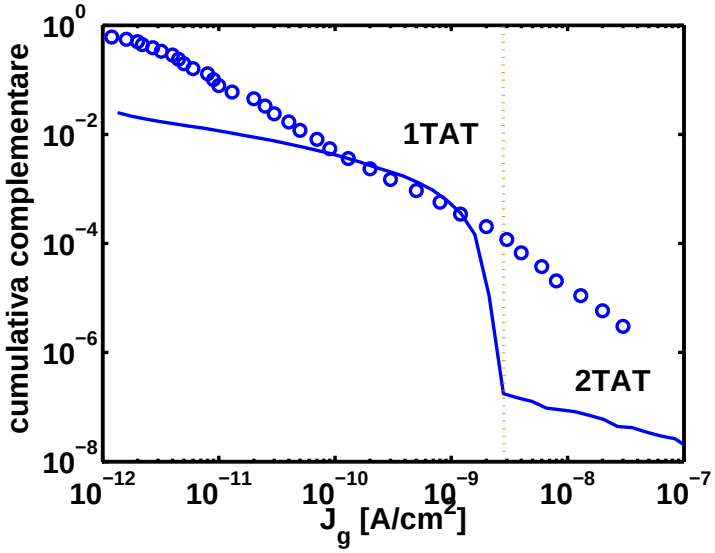


Figura 3.11: *Curva sperimentale (pallini) e simulata (linea continua) per $t_{ox} = 6.5nm$ con generazione di trappole non correlata. Sono indicate i tratti di curva relativi alla conduzione 1TAT e 2TAT.*

ciascuna di esse.
Ovviamente il metodo può essere reso più complesso andando a considerare separatamente $X = 2$, $X = 3$, ecc., modificando di conseguenza i pesi associati ai vari casi e riuscendo, potenzialmente, a raggiungere valori di densità di probabilità ancora inferiori.

Prima di procedere oltre, esaminiamo brevemente perché invece il metodo dello *Splitting* non si è rivelato adatto al nostro caso. In sostanza, siamo in un contesto probabilistico in cui sono di difficile individuazione dei livelli e sottolivelli in cui dividere lo spazio degli eventi, non esistono "traiettorie" da controllare, per poi fermarle o moltiplicarle (vedi Par. 3.3.3). E questo avviene perché l'assegnazione dei parametri statistici avviene, per così dire, in "parallelo" (all'inizio) e non in "serie" (ricorsivamente all'interno della stessa simulazione). Per questi motivi la *Splitting Tecnique* risulta completamente inefficace.

Se quindi la tecnica dell'*Importance Sampling* consente di raggiungere le densità di probabilità desiderate, c'è però da fare un'ultima considerazione per quanto riguarda l'interpolazione dei dati sperimentali. In Fig. 3.11 è mostrata la curva relativa ai dati sperimentali con $t_{ox} = 6.5nm$ e la curva simulata. È chiaramente visibile un crollo di quest'ultima per densità di correnti attorno ai $10^{-9} A/cm^2$. Il tratto di distribuzione precedente è imputabile alla conduzione di tipo 1TAT, quello successivo alla conduzione di tipo 2TAT. Il motivo di tale crollo è da ricercarsi nel fatto che esiste un determinato valore dei parametri

di conduzione 1TAT in grado di massimizzare il processo, ed oltre tale valore di corrente ci si può arrivare solo con meccanismi di conduzione superiori, per esempio il 2TAT. L'entità del crollo, invece, è determinata dal numero di realizzazioni in cui i difetti effettivamente cooperano per dare vita (almeno) al 2TAT. Evidentemente, nel nostro modello simulativo ciò ciò avviene con probabilità molto minore di quanto non accada in realtà, come dire che se anche il numero di trappole per cella è ≥ 2, è molto raro osservare una loro interazione. È molto probabile invece che ognuna agisca in maniera indipendente, creando così molteplici percorsi conduttivi 1TAT. Questo a sua volta è dovuto al fatto che, essendo le dimensioni del dispositivo y e z molto maggiori dello spessore dell'ossido di tunnel, è molto improbabile che due o più trappole si "vedano", a causa della grande distanza trasversale che le separa.

L'unica via d'uscita a questo problema è stata quella di considerare la generazione di trappole correlata, formalizzata nel Par. 3.2.4. In questo modo è molto più probabile che i difetti cooperino, il che vuol dire che il crollo mostrato dalla curva simulata viene colmato in maniera tanto maggiore quanto più ravvicinati vengono creati, almeno ad un'analisi del primo ordine. La prima trappola viene generata secondo la statistica originaria, mentre le successive correlate con i seguenti parametri: funzione di probabilità riferita alla singola trappola di tipo gaussiano, con deviazione standard $R_{corr} = 0.125nm$ e utilizzo del metodo semplificato del baricentro. Il numero di trappole rimane conforme alla statistica di Poisson, solo il loro posizionamento spaziale viene modificato. Il valore estremamente basso del raggio di correlazione R_{corr} è da ritenersi conseguenza della carica elettrica posseduta dai difetti stessi, che se per grandi distanze può essere scaricata sulla *cross section* efficacie aumentandone il valore, nel caso di brevi scostamenti ciò non risulta più essere sufficiente a causa dell'interazione ancora maggiore tra i due potenziali coulombiani. R_{corr} è quindi da intendersi come una sorta di distanza efficace, più che come distanza reale.

Il risultato finale è mostrato in Fig. 3.12, questa volta per tutti e tre gli spessori del *tunnel oxide*. Come si può constatare, per tutti si è raggiunto un buon *fitting*, tralasciando la parte iniziale di ciascuna curva, derivante dalla statistica sul *tunneling* FN, che, come dicevamo, nel nostro modello è concepito di natura deterministica. La densità di difetti per unità di volume è $N_T = 10^{15} cm^{-3}$ per tutti gli spessori. Solo per $t_{ox} = 6.5nm$ è stata effettuata una simulazione che arrivasse a densità di probabilità dell'ordine di 10^{-9}, perchè nonostante il metodo di *enhancement* statistico utilizzato si è trattato comunque di far fronte ad un numero di iterazioni totali molto ingente (dell'ordine di $5 \cdot 10^5$), che comincia ad essere proibitivo per i normali tempi di laboratorio. Sono ancora visibili le porzioni di curva relative al 1TAT e al 2TAT, che questa volta formano pressoché un continuo. La coda 2TAT, per il caso $t_{ox} = 6.5nm$, mostra un leggero cambio di pendenza rispetto a quella del

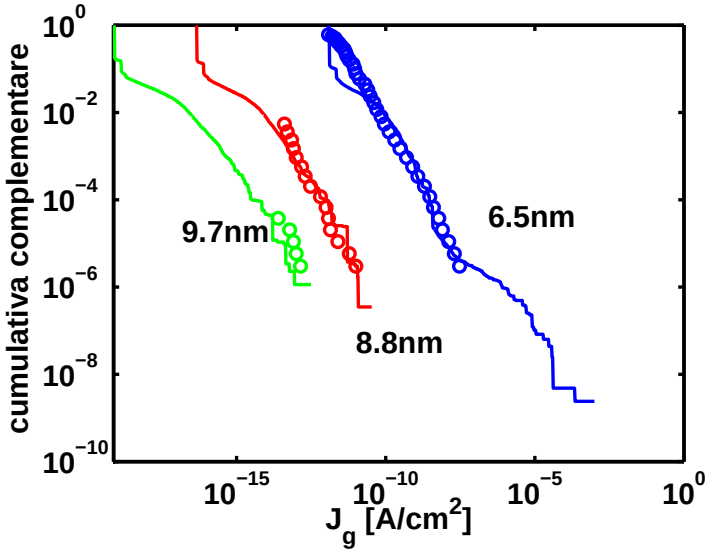


Figura 3.12: *Curve sperimentali (pallini) e curve simulate (linee continue) relative ai tre spessori. La generazione di trappole è, questa volta, correlata.*

1TAT, tuttavia in assenza di dati sperimentali non possiamo affermare se sia o meno verosimile.

§3.5 Conclusioni

In questo capitolo è stata fornita una panoramica sui metodi di generazione di variabili casuali adottati nel corso delle simulazioni tridimensionali, variabili che rappresentano parametri fondamentali delle trappole, la cui natura è prettamente aleatoria, cercando di fornire per ognuno di essi una derivazione matematica sufficientemente rigorosa. Successivamente si sono introdotti in forma generale i metodi di *enhancement* statistico per simulazioni Monte Carlo, simulazioni che hanno permesso di determinare, nel caso dell'*array* di memoria FLASH sotto esame, la distribuzione spaziale ed energetica delle trappole all'interno del *tunnel oxide* e di constatare come le suddette trappole siano spazialmente correlate. L'efficacia del metodo *Importance Sampling* è evidente qualora si consideri che, per quanto riguarda la determinazione della distribuzione cumulativa delle correnti di *leakage*, probabilià dell'ordine di 10^{-9} sono state raggiunte con un numero di iterazioni di circa $5 \cdot 10^5$.

Capitolo 4

Affidabilità di memorie FLASH in ritenzione

An expert is a person who has made all the mistakes that can be made in a very narrow field.

(L'esperto è una persona che ha fatto in un campo molto ristretto tutti i possibili errori.)

Niels Bohr

In questo capitolo verrà effettuata un'analisi dell'affidabilità in ritenzione per quanto riguarda un array *di memorie FLASH di tipo NOR multilivello. In particolare, verrà individuata la densità massima di difetti accettabile nel* tunnel oxide*, funzione dello spessore e dei criteri di fallimento, oltre la quale il banco di memoria non è più cosiderato affidabile.*

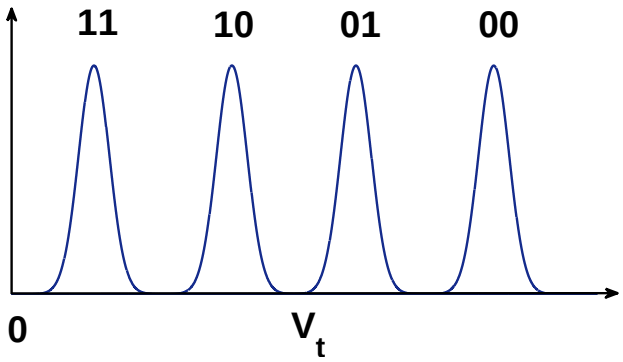


Figura 4.1: *Esempio di distribuzione delle tensioni di soglia corrispondenti ai diversi livelli logici di una cella FLASH NOR multilivello a 2 bit.*

§4.1 Introduzione

Nel corso degli ultimi anni, le memorie FLASH hanno dovuto far fronte ai grossi problemi derivanti dallo *scaling* delle celle. Uno dei metodi per "aggirare" questi limiti, in grado di raggiungere così maggiori densità di informazione a parità di area di *chip*, è l'utilizzo di celle di tipo *multilivello*. Originariamente, infatti, la cella di memoria FLASH era in grado di immagazzinare l'informazione contenuta in un solo bit, dal momento che erano possibili due soli stati fisici distinti corrispondenti uno allo 0 logico e l'altro al 1 logico. Con l'affinarsi della tecnologia, però, si è riusciti ad aumentare in maniera sufficientemente affidabile il numero di stati fisici possibili, in modo da poter stoccare un numero di bit maggiore di uno. In particolare, se con n indichiamo il numero di bit, il numero dei livelli logici è 2^n. In Fig. 4.1 è mostrato un esempio di cella multilivello a 2 bit, con i relativi livelli logici e le distribuzioni delle tensioni di soglia.

Al crescere del numero di livelli, però, i requisiti di affidabilità diventano sempre più stringenti. Rispetto alla cella a singolo bit, infatti, il *range* di tensioni utilizzate deve essere allargato, i margini di lettura sono ridotti, così come devono essere ridotte le imperfezioni circuitali e le distribuzioni delle tensioni di soglia devono essere più sottili. Queste problematiche comportano l'utilizzo di campi elettrici più cospicui in fase di programmazione, cancellazione e lettura, minando l'affidabilità del dispositivo stesso. Non di meno, la velocità di

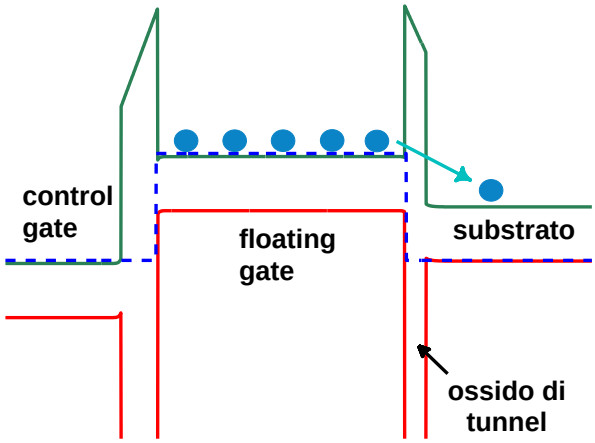


Figura 4.2: *Struttura a bande di una cella FLASH NOR in ritenzione. Dal floating gate gli elettroni tendono a rilassare verso il substrato, che si trova a più bassa energia. In questo modo la tensione di soglia del dispositivo varia in ragione della quantità di carica persa. In blu tratteggiato sono indicati i livelli di Fermi nelle tre zone di semiconduttore.*

lettura e di scrittura diminuisce, mentre i potenziali disturbi aumentano [41].

In particolare, andiamo ora ad approfondire il concetto affidabiltà in ritenzione. Innanzi tutto, la ritenzione consiste nella capacità della cella di conservare il dato memorizzato. Come abbiamo visto nel Cap. 1, la memorizzazione del dato corrisponde ad un immagazzinamento di carica nella regione di *floating gate*, che a sua volta provoca uno *shift* della tensione di soglia del dispositivo. Questa carica, tuttavia, non è confinata nel FG in maniera così impeccabile: tende a sfuggire verso il substrato attraversando la barriera energetica mostrata dal *tunnel oxide* con i meccanismi di conduzione studiati nei capitoli precedenti, ovvero il *tunneling* Fowler–Nordheim e il TAT. In Fig. 4.2 è mostrata la struttura a bande di una cella FLASH NOR in ritenzione. Si può osservare il *floating gate* carico di elettroni e perciò ad elevata energia mentre il substrato ed il *control gate* sono al potenziale di riferimento (osservare i livelli di Fermi). Gli elettroni confinati tenderanno quindi a ripristinare l'equilibrio termodinamico rilassando più verso il substrato che non verso l'elettrodo di *gate*, dal momento che il *control oxide* è, tipicamente, più spesso del *tunnel oxide*. Man mano che la carica viene persa, la tensione di soglia varia in maniera corrispondente fino ad arrivare, per tempi sufficientemente lunghi, ad un valore significativamente diverso da quello originale. Se lo *shift* oltrepassa un certo limite, si parla quindi di fallimento in ritenzione. Specifiche commerciali

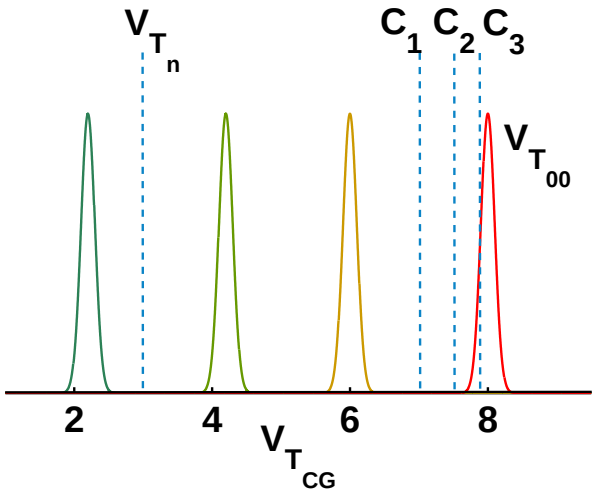


Figura 4.3: *Rappresentazione qualitativa delle distribuzioni delle tensioni di soglia per la memoria FLASH NOR sotto esame. Sono indicate la soglia neutra a 3V e le tensioni critiche relative ai criteri di fallimento* $C_1 = 7V$, $C_2 = 7.5V$ *e* $C_3 = 7.9V$.

richiedono un'affidabilità di almeno dieci anni.

Si è deciso, quindi, di studiare quantitativamente questo fenomeno nel caso di una memoria FLASH di tipo NOR multilivello a due bit, sfruttando gli algoritmi di calcolo delle correnti di perdita sviluppati sinora nonché delle tecniche di *enhancement* statistico descritte in dettaglio nel Cap. 3. Le caratteristiche della memoria in esame sono: celle di tipo nMOS con tensione di soglia neutra $V_{T,n} \approx 3V$, tensioni di soglia programmate $V_{T,11} = 2.2V$, $V_{T,10} = 4.2V$, $V_{T,01} = 6V$, $V_{T,00} = 8V$ e area attiva $L \cdot W = 110nm \times 40nm$.

§4.2 Densità di difetti e spessore del tunnel oxide

La corrente TAT è il meccanismo di conduzione più critico per quanto riguarda la migrazione degli elettroni dal *floating gate* verso il substrato, dal momento che, almeno per le architetture NOR, spessori del *tunnel oxide* dell'ordine di $9 - 10nm$ sono più che sufficienti a garantire una ritenzione di almeno dieci anni nel caso che la corrente di perdita fosse dovuta solamente al *tunneling* diretto/Fowler–Nordheim. Tuttavia, come abbiamo spiegato nel capitolo precedente, i parametri relativi alle trappole sono, per dispositivi così piccoli, governati da leggi statistiche e uno studio del fenomeno basato sull'utilizzo dei soli valori medi non risulta più sufficientemente accurato. È

$t_{ox}[nm]$	$V_{T_{FG}}[V]$	α_G	$C_{CG}[\mu F/cm^2]$	$Q_{FG_0}[\mu C/cm^2]$	$V_{FG_0}[V]$
10	2	0.67	0.9	-4.5	-3.33
9.5	1.9	0.63	0.84	-4.2	-3.16
9	1.8	0.6	0.81	-4.05	-3
8.5	1.7	0.57	0.79	-3.95	-2.83
8	1.6	0.53	0.78	-3.75	-2.56
7.5	1.5	0.5	0.78	-3.75	-2.4
7	1.4	0.47	0.78	-3.75	-2.24

Tabella 4.1: *Parametri relativi alla cella al variare dello spessore del tunnel oxide, mantenendo fissate le specifiche riguardo alle tensioni di soglia.*

neccessario allora affrontare anche il problema della ritenzione in maniera statistica, andando a considerare la distribuzione delle tensioni di soglia al variare del tempo e valutando la percentuale di celle che risultano inaffidabili secondo un certo criterio di fallimento. Lo studio è stato effettuato per il caso più critico, ovvero quando $V_T = V_{T,00}$ (tensione di soglia massima, quantità di carica immagazzinata nel *floating gate* massima, tensione dello stesso il più negativa possibile). In ritenzione, il substrato ed il *control gate* sono entrambi al potenziale di riferimento, *i.e.* $0V$. Lo scopo, innanzi tutto, è di stabilire una relazione tra spessore dell'ossido di tunnel e densità di difetti critica per unità di volume all'interno dello stesso. Risulta intuitivo, infatti, che diminuendo lo spessore anche la concentrazione critica diminuisca, non tanto per il crescere della corrente di *tunneling* tradizionale che va a sommarsi a quelle TAT e che è ancora trascurabile, ma quanto per il fatto che le trappole stesse vedono una barriera di potenziale più sottile e quindi più trasparente, rendendo più copiosa la perdita. Ne sono quindi sufficienti un numero inferiore per causare lo stesso flusso di carica. Per prima cosa, valutiamo come si modificano i parametri della cella al variare dello spessore dell'ossido, mantenendo invariate le richieste sulle tensioni di soglia esplicitate in precedenza. Assumiamo che, per il valore standard $t_{ox} = 10nm$, il coefficiente di accoppiamento valga $\alpha_G = 2/3$, valore tipico per questa tecnologia. Dall'Eq. 1.9, quindi, $V_{T_{FG}} = 2V$, e supponiamo che valga l'approssimazione $V_{T_{FG}} = 2\Phi_B + \frac{Q_{channel}}{C_{ox}} + V_{FB} \approx \frac{Q_{channel}}{C_{ox}}$, in modo tale che $V_{T_{FG}} \propto tox$. Ammettiamo, secondo l'Eq. 1.6, che $\alpha_G = \frac{C_{CG}}{C_{CG} + C_{ox}}$, dal momento che il substrato è in accumulo di lacune e che possiamo trascurare i contributi di *source* e *drain*. Conoscendo C_{CG} possiamo ricavare la carica iniettata nel *floating gate* in fase di programmazione Q_{FG_0}[1] sempre attraver-

[1] considerando che lo *shift* della tensione di soglia tra lo stato neutro e lo stato "00" è di $8 - 3 = 5V$.

so l'Eq. 1.9, e arrivare a determinare il valore della tensione V_{FG_0} tramite l'Eq. 1.8. Nella Tab. 4.1 sono riassunti i parametri della cella al variare dello spessore dell'ossido, parametri utilizzati per la determinazione dello *shift* della tensione di soglia $V_{T,00}$ nel corso delle simulazioni. Si sono considerati valori di t_{ox} da un valore standard di $10nm$ fino ad un valore minimo di $7nm$.

In secondo luogo, bisogna determinare un criterio di fallimento. Si è scelto, quindi, come densità di difetti critica, quella concentrazione tale per cui la distribuzione cumulativa raggiunge un valore limite di 10^{-9} con uno *shift* di soglia di $0.1V$, $0.5V$ e $1V$ rispettivamente. In Fig. 4.3 sono mostrate qualitativamente le distribuzioni delle tensioni di soglia, i valori della soglia neutra e le tensioni "critiche" di $7V$, $7.5V$ e $7.9V$. Il valore di $7V$ corrisponde ad uno *shift* di $1V$, che è veramente il massimo scostamento ancora tollerabile: la distanza tra $V_{T,00}$ e $V_{T,01}$ è, infatti, di circa $2V$, indi percui si assegna, al massimo, il 50% dell'intervallo a ciascuno dei due livelli.

Terzo, si è scelto come modello di conduzione il modello NTAT tridimensionale descritto nel Cap. 2, mentre, per quanto riguarda la tecnica Monte Carlo utilizzata, si fa sempre uso dell'*Importance Sampling* descritto nel Cap. 3. L'unica variazione rispetto ad allora sta nel fatto che Ψ rappresenta ora l'evento: "lo scostamento della tensione di soglia della cella dallo stato programmato equivale ad un certo valore". Ovviamente, a questo punto, Ψ dipende anche dal tempo di ritenzione, oltre che dalla qualità delle trappole che generano il TAT. Per queste ultime, infine, si è scelto come parametri statistici gli stessi risultanti dall'analisi dell'*array* di memorie del capitolo precedente, dal momento che la tecnologia e gli spessori dei dielettrici sono molto simili:

- *Cross section* e tempo caratteristico delle trappole: $\sigma = 4 \cdot 10^{-10} cm^2$ e $\tau_{att} = 10^{-15} s$.
- Numero di trappole per cella che segue la distribuzione di Poisson (vedi Par. 3.2.3), di valor medio $N = N_T \cdot Vol_{ox}$, dove con N_T si è indicata la densità media di trappole per unità di volume e con Vol_x il volume del *tunnel oxide*. N_T è un parametro della simulazione, così come lo è Vol_x.
- Distribuzione spaziale uniforme, sia lungo la direzione longitudinale x che sulle trasversali y e z.
- Distribuzione energetica di tipo gaussiano (vedi Par. 3.2.2), con valor medio $\bar{E} = -1.2eV$ e deviazione standard $\sigma_E = 0.5eV$.
- Generazione di trappole correlata[2], secondo una densità di probabilità gaussiana a simmetria sferica, con deviazione standard $R_{corr} = 0.125nm$.

[2]In merito alle considerazioni sui valori delle *cross section* e della distanza di correlazione, si rimanda il lettore alla Sez. 3.4.

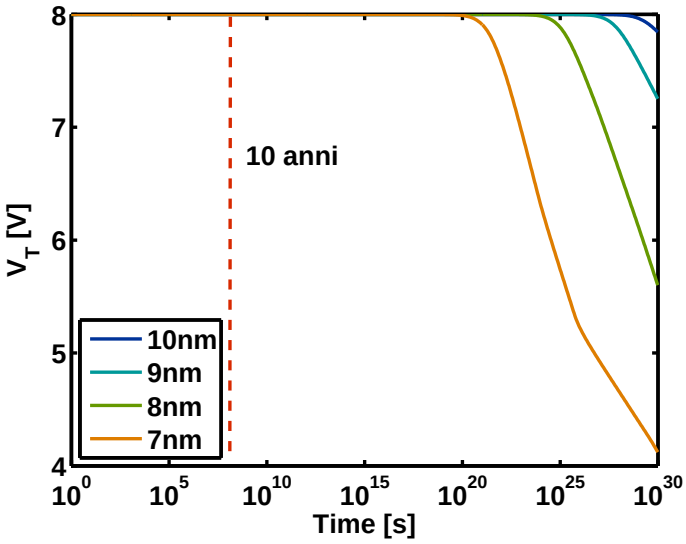


Figura 4.4: *Contributo alla diminuzione della tensione di soglia della sola corrente di tunneling diretto/Fowler–Nordheim, per i quattro spessori di ossido indicati.*

Entriamo ora nel merito delle simulazioni. Innanzi tutto si è proceduto andando a valutare l'ammontare della carica trasferita dal *floating gate* al substrato in un certo lasso di tempo, per poi sottrarre questa alla carica immagazzinata inizialmente Q_{FG_0}. Il valore risultante permette di andare ad aggiornare sia quello della tensione di FG, tramite l'Eq. 1.8, sia, cosa piú importante, quello della tensione di soglia al *control gate*, secondo l'Eq. 1.1. L'operazione si è ripetuta fintantoché non si è giunti ad un tempo totale prefissato, campionando gli intervalli logaritmicamente, data l'estensione dell'intervallo temporale complessivo stesso.

Valutiamo, preliminarmente, il contributo alla diminuzione della tensione di soglia della sola corrente di *tunneling* diretto/Fowler–Nordheim, fenomeno intrinseco ineliminabile. Come si vede in Fig. 4.4, anche nel caso del minimo spessore di ossido occorre aspettare all'incirca $10^{20}s$ per ottenere uno scostamento significativo dal valore programmato, un tempo sufficientemente lungo per qualsiasi tipo di applicazione[3], considerato che dieci anni corrispondono a $3.15 \cdot 10^8 s$. Ciò significa che i problemi legati alla ritenzione, qualora ve ne fossero, sono causati dalla corrente TAT. Nella Fig. 4.5 sono, infatti, mostrati degli esempi di realizzazioni IV proprio di corrente TAT: la caratteristica che limita inferiormente il fascio di curve corrisponde alla corrente di tunnel, che "assorbe" le eventuali TAT poco efficienti, mentre al di sopra di esso si possono apprezzare quelle che invece danno contributi via via più considerevoli. La

[3] $10^{20}s$ =mille miliardi di anni.

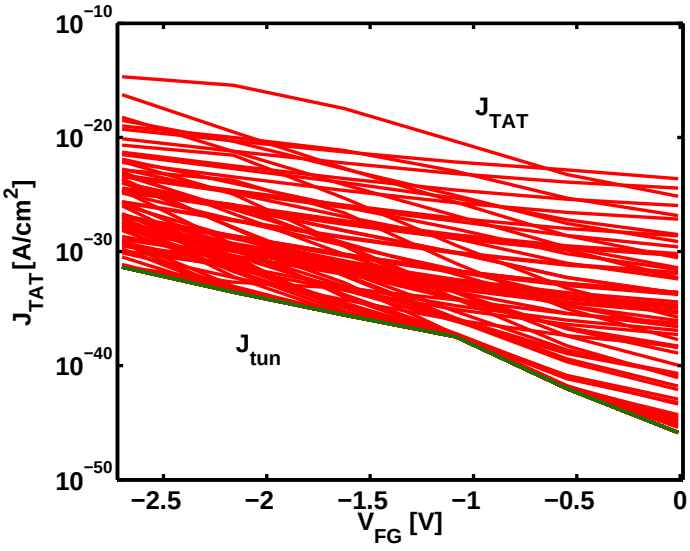
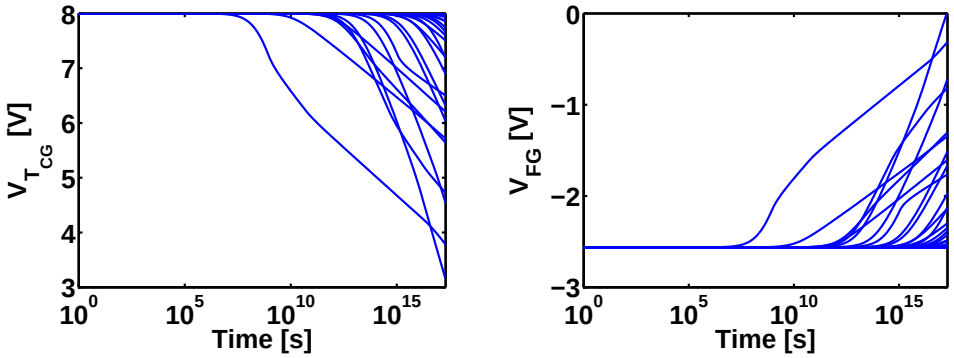


Figura 4.5: *Esempi di realizzazioni per la corrente di TAT per $t_{ox} = 9nm$: la curva verde rappresenta la corrente di tunneling tradizionale, le curve rosse invece le correnti TAT più efficaci.*

figura fa riferimento a $t_{ox} = 9nm$, con in ascissa la tensione dell'elettrodo di *floating gate*, negativa dal momento che la carica immagazzinata è costituita da elettroni, e compresa in un *range* che va circa dal minimo valore che l'elettrodo assume V_{FG_0}, conseguente alla programmazione, fino alla tensione nulla, che raggiunge nella malaugurata ipotesi in cui tutta la carica del FG venga persa. In Fig. 4.6(a) sono mostrati i rispettivi e più significativi andamenti della tensione di soglia in funzione del tempo, mentre in Fig. 4.6(b) sono rappresentate le tensioni di *floating gate*. Così come la V_{FG} tende al valore d'equilibrio, $0V$, così la $V_{T_{CG}}$ tende al suo valore neutro, in questo caso $V_{T_{CG,n}} = 3V$.
Passiamo ora ad una visione statistica. Nelle Figs. dalla 4.7(a) alla 4.7(d) sono mostrate[4] le distribuzioni cumulative relative allo *shift* della tensione di soglia a partire dal valore critico $V_{T,00} = 8V$, calcolate dopo un tempo di ritenzione di dieci anni e per spessori del *tunnel oxide* di $t_{ox} = 7, 8, 9, 10nm$. In ogni grafico sono mostrate diverse curve: ciascuna fa riferimento ad una differente densità di difetti, in questo modo è possibile risalire alla concentrazione critica mediante interpolazione (su scala logaritmica) al variare del criterio di fallimento.

Osservando le figure, appare evidente come, al crescere dello spessore dell'ossido e a pari densità, la distribuzione si restringe verso il valore di programmazione di $8V$ e il, diciamo così, punto di attacco della coda di probabilità scende verso valori inferiori. Infatti, nonostante il numero medio di trappole

[4]Non mostrate, sono state eseguite simulazioni anche per $t_{ox} = 7.5, 8.5, 8.75, 9.5, 9.75nm$.



(a) (b)

Figura 4.6: *Andamento nel tempo delle tensioni di soglia (a) e di floating gate (b), relative alle correnti rappresentate in Fig. 4.5.*

per cella cresca al crescere dello spessore, sempre a pari densità, è altresì vero che, tra esse, un maggior numero risultano poco efficienti ai fini della conduzione di corrente, proprio in virtù dell'incrementato spessore di barriera. Quindi, pur essendoci (mediamente) più difetti, sono comunque meno quelli che danno contributi "visibili". Il decremento del valore di probabilità del punto di attacco è quindi giustificato.

Un'altra caratteristica risalta immediatamente: a partire dallo spessore di $8nm$, verso *shift* di soglia elevati, si può notare un repentino crollo della distribuzione. Questo è conseguenza del fatto che la porzione di curva che va dal punto di attacco al punto di crollo è da imputare a fenomeni di conduzione 1TAT, mentre la porzione successiva alla conduzione 2TAT. Prendiamo in considerazione il trasporto mediato da una singola trappola: esiste un valore dei parametri relativi ad essa in grado di massimizzare il meccanismo di conduzione, fornendo così una sorta di "estremo superiore" per la corrente 1TAT. Solo un meccanismo di conduzione a due trappole può oltrepassare questo limite, fintantoché anche per questo non si raggiungano i parametri che massimizzano il processo, dopodiché occorre aumentare il numero di difetti cooperanti a tre, e così via (analogamente al caso studiato nella Sez. 3.4). Il crollo visibile nei grafici corrisponde appunto al limite 1TAT–2TAT. Come si vede in maniera chiara nel grafico relativo allo spessore di $9nm$, l'entità del salto diminuisce al crescere della densità: maggiore è il numero di difetti, maggiore à la probabilità di avere il 2TAT rispetto al 1TAT. Nel Cap. 4 abbiamo utilizzato la generazione correlata di difetti proprio allo scopo di appianare questo salto, non riscontrato nelle misure sperimentali. Il motivo per cui ora, pur utilizzando ancora la stessa generazione correlata, il salto è chiaramente visibile, è che in quel contesto si aveva a che fare con densità dell'ordine di $10^{15}cm^{-3}$,

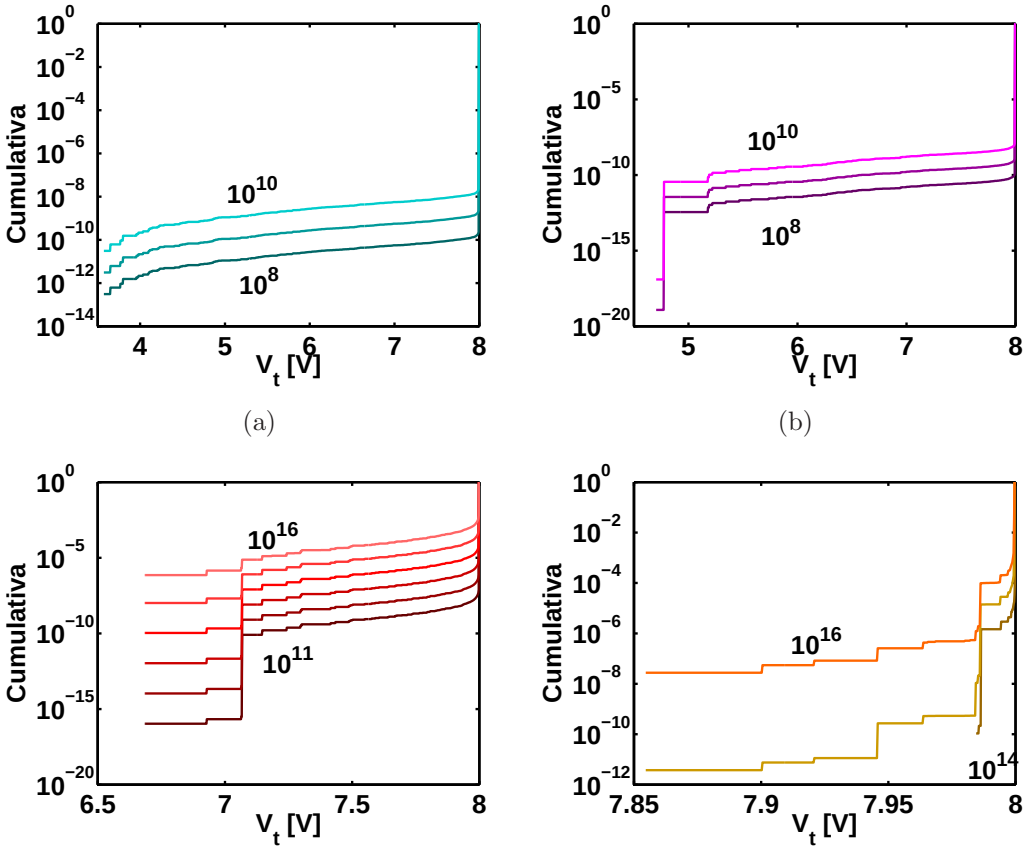


Figura 4.7: *Distribuzioni cumulative della tensione di soglia dopo un tempo di dieci anni in condizioni di ritenzione. Il valore originale è $V_{T,00} = 8V$. In (a) abbiamo $t_{ox} = 7nm$, in (b) $t_{ox} = 8nm$,, in (c) $t_{ox} = 9nm$, e in (d) $t_{ox} = 10nm$. All'interno di ogni grafico sono indicate i valori minimi e massimi della concentrazione di difetti, simulate di decade in decade tra i suddetti valori.*

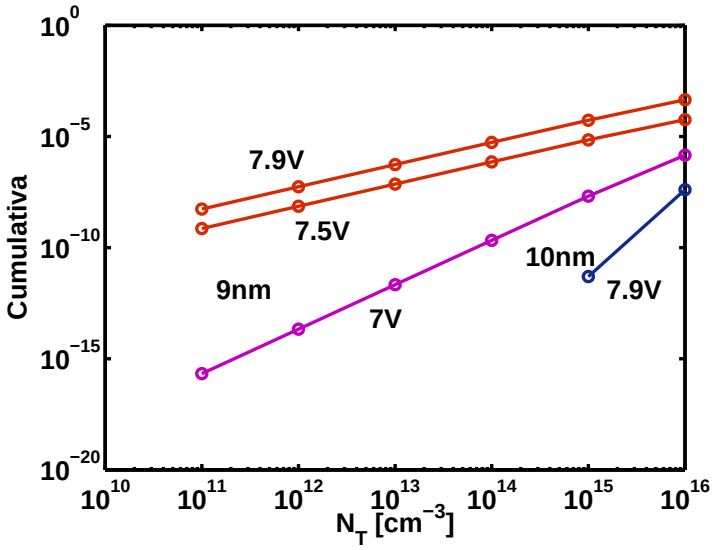


Figura 4.8: *Valore della distribuzione cumulativa alle tensioni critiche di 7V, 7.5V e 7.9V per $t_{ox} = 9nm$ e di 7.9 per $t_{ox} = 10nm$. Le curve rosse hanno pendenza 1, la violetto 2 e la blu 4, corrispondenti ai regimi di conduzione 1, 2 e 4TAT.*

mentre ora stiamo considerando soprattutto concentrazioni di valore inferiore. Nondimeno, nei casi in cui si è simulata una densità di tal valore, come in $t_{ox} = 9nm$ e $N_T = 10^{15} \div 10^{16}$, si nota come il salto sia di lieve entità e come la coda 2TAT tenda a formare un continuo con quella 1TAT. È interessante constatare che, per quanto riguarda lo spessore di $10nm$, si sia in presenza di più di un crollo: al valore critico di $7.9V$ ci si arriva addirittura con la coda 4TAT.

Un'ulteriore considerazione merita di essere fatta per quanto riguarda la dipendenza delle distribuzioni dalla densità di difetti a parità di spessore dell'ossido. Come si affermava in precedenza, a governare la quantità di trappole che danno contributi visibili allo *shift* di soglia è, in massima parte, lo spessore stesso. Pertanto, anche variando la densità di ordini di grandezza, la percentuale di trappole sensibili rispetto al totale non cambia. Questo ragionamento, valido se la concentrazione è sufficientemente bassa da garantire che $\Pr(N+1) \ll \Pr(N)$, $\forall N$, dove N è il numero di trappole, ha come conseguenza che le distribuzioni cumulative, a parità di spessore dell'ossido, al variare della densità di trappole semplicemente traslano rigidamente in senso verticale. Infatti, la coda 1TAT è legata alla probabilità di avere una sola trappola per cella, dato che i contributi 1TAT dovuti a più trappole per cella sono di probabilità molto inferiore. Quindi, secondo Poisson, $\Pr(1) = e^{-\lambda}\lambda$, con $\lambda = N_T \cdot Vol_{ox}$, ma essendo $\lambda \ll 1$, $\Pr(1)$ può essere scritta

$$\begin{aligned}\Pr(1) &= e^{-\lambda}\lambda \\ &\approx (1-\lambda)\lambda \\ &\approx \lambda\end{aligned}$$

dove nel primo passaggio si è espanso l'esponenziale nella serie di McLaurin arrestata al primo grado. Quindi, ad una variazione di una decade della densità di difetti, corrisponde una variazione di una decade anche per il valore di probabilità della coda 1TAT.
Lo stesso ragionamento può essere applicato alla coda 2TAT

$$\begin{aligned}\Pr(1) &= \frac{e^{-\lambda}\lambda^2}{2} \\ &\approx \frac{(1-\lambda)\lambda^2}{2} \\ &\approx \frac{\lambda^2}{2} \\ &\propto \lambda^2\end{aligned}$$

da cui si evince come, per una variazione di una decade della densità di difetti, corrisponda una variazione di due decadi per la coda di probabilità 2TAT. Il discorso è analogo anche per le code 3 e 4TAT. Dei risultati quantitativi sono riportati in Fig. 4.8, dove sono mostrati gli andamenti della distribuzione cumulativa alle tensioni critiche, al variare della densità di difetti. Per $t_{ox} = 9nm$ a $7.9V$ e a $7.5V$ si è in regime 1TAT, e difatti la pendenza è unitaria (in scala log–log), mentre per $7V$ siamo in regime 2TAT, e la pendenza raddoppia. Per $t_{ox} = 10nm$, l'unico dato disponibile è quello a $7.9V$ e mostra una pendenza quadrupla, essendo in regime 4TAT.

Attraverso interpolazioni di questo tipo, eseguite per ogni spessore simulato, è possibile arrivare alla determinazione della concentrazione di difetti critica rispetto ai tre diversi criteri di fallimento. In Fig. 4.9 sono mostrati appunto gli andamenti della densità critica in funzione dello spessore del *tunnel oxide*. A seconda dello spessore dell'ossido, si possono individuare delle diverse dipendenze funzionali. Per ossidi via via sempre più sottili, infatti, la densità critica decresce sempre più blandamente, quasi in maniera asintotica verso valori dell'ordine di $10^8 cm^{-3}$, in modo quasi indipendente dal criterio di fallimento scelto. Questo è, d'altra parte, ragionevole, se si pensa che con l'assottigliarsi dello strato di isolante quasi tutte le trappole, quando esistono, sono in grado di dare un contributo significativo allo *shift* della tensione di

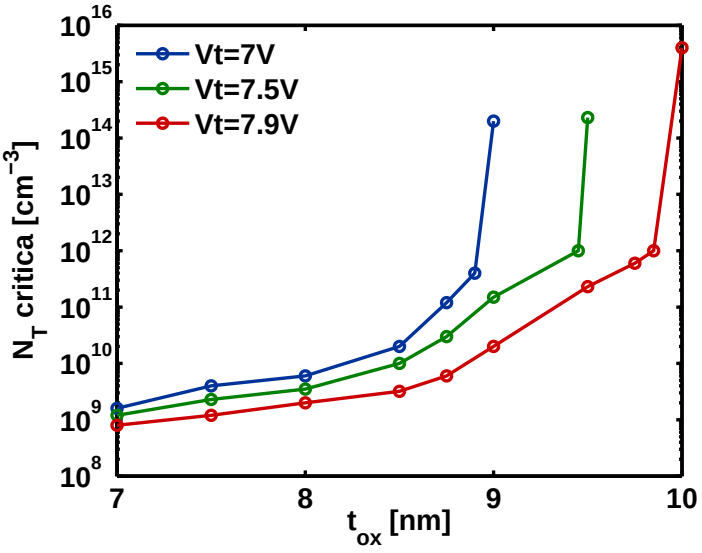


Figura 4.9: *Valore della densità critica di difetti calcolato in funzione dello spessore del tunnel oxide, per una probabilità di fallimento di* 10^{-9} *alle tensioni di 7V, 7.5V e 7.9V.*

soglia. Come dire, non ci sono più o sono veramente poche le trappole "invisibili" ai fini della statistica delle correnti di *leakage*. Infatti, la distribuzione cumulativa, per tali spessori, si allunga verso tensioni sempre più basse e al contempo si appiattisce verso il punto di attacco. Portando il ragionamento al limite, il valore asintotico raggiunto dalla densità critica potrebbe essere calcolato come la probabilità di avere una trappola ogni 10^9 celle, dal momento che quell'unica cella difettosa risulterebbe fallimentare. E questo in maniera quasi indipendente, in linea di principio, dal criterio di fallimento. In formule, questo concetto è esprimibile come la probabilità di avere nella singola cella un numero di trappole maggiore o uguale ad uno del valore di 10^{-9}

$$\begin{aligned}\Pr(N \geq 1) &= 1 - \Pr(0) \\ &= 1 - e^{-\lambda} \\ &\approx 1 - (1 - \lambda) \\ &\approx \lambda \\ &= 10^{-9}\end{aligned}$$

sempre sfruttando lo sviluppo in serie di McLaurin dell'esponenziale, dal momento che $\lambda \ll 1$. Considerando che appunto $\lambda = N_T \cdot Vol_{ox}$, basta invertire l'equazione per giungere ad un valore di concentrazione critica $N_T \sim 5 \cdot 10^7 cm^{-3}$, per un ossido da $7nm$. Ovviamente questo valore teorico non è veramente

asintotico, dal momento che dipende da t_{ox}. Tuttavia la dipendenza è lineare, quasi irrilevante in un grafico bilogaritmico come quello in Fig. 4.9. Il valore estrapolato dalle simulazioni è invece sempre leggermente superiore a quello teorico per il fatto che esistono ancora delle trappole poco efficienti ai fini della conduzione, il cui contributo rimane pressoché irrilevante, perciò il numero medio deve essere lievemente più grande per compensare questa perdita.

Per ossidi sempre più spessi, invece, accade un altro fenomeno curioso, che questa volta fa impennare la densità critica, ed è, in aggiunta, sostanzialmente dipendente dal criterio di fallimento richiesto. Fissato quest'ultimo, infatti, e aumentando lo spessore di isolante, come dicevamo in precedenza la distribuzione cumulativa si restringe verso il valore iniziale della tensione di soglia, fintanto che il punto di fallimento non è più intersecato dalla coda 1TAT, ma è necessario passare a quella relativa al 2TAT. In parole povere, anche la più efficiente delle trappole, per quanto riguarda il 1TAT, non è in grado di far fallire una cella secondo i requisiti prefissati. Occorre che il meccanismo di conduzione sia più efficace, passando quindi al TAT di ordine superiore. Nel far questo, però, bisogna oltrepassare il punto di crollo della distribuzione: essendo la probabilità di fallimento scesa repentinamente, per riportarla a valori dell'ordine dei precedenti è necessario aumentare in maniera drastica il valore della densità di difetti. Dopo questo breve intervallo di inflazione, la concentrazione critica in funzione dello spessore torna a salire con pendenze più dolci, fino a che anche la coda 2TAT non si esaurisce, passando ad un regime 3TAT che comporta un nuovo balzo della densità, ecc. In Fig. 4.9 sono mostrati i balzi relativi al passaggio dal 1TAT al 2TAT, per tensioni critiche di $7V$ e $7.5V$. Non si è, purtroppo, riusciti a procedere oltre a causa della scarsa affidabilità delle simulazioni in quanto molto onerose dal punto di vista computazionale. Per la tensione critica di $7.9V$, invece, il salto rappresentato è comprensivo dei passaggi dal 1TAT al 4TAT, così ravvicinati in termini di spessore dell'ossido che è ben difficile riuscire a discernerli singolarmente. Per questo, l'entità dell'incremento è maggiore che non negli altri due casi. In Fig. 4.10, infine, è mostrata l'estrapolazione delle tensioni, chiamiamole di "confine", che individuano il crollo della distribuzione cumulativa dovuto al passaggio dal regime 1TAT al 2TAT e dal 2TAT al 3TAT, rispettivamente, al variare dello spessore dell'ossido. La seconda è stata ricavata considerando la minima tensione raggiunta dalla coda 2TAT, al di là della quale si suppone si verifichi il crollo e il conseguente cambio di regime di conduzione. Entrambe tendono, teoricamente, al valore asintotico di $8V$, per il fenomeno di "schiacciamento" della cumulativa verso l'originale $V_{T,00}$. Confrontando questo grafico con quello in Fig. 4.9 si può constatare la coerenza tra i valori dello spessore dell'ossido per i quali si ha il balzo della densità di trappole critica a fissata tensione di fallimento con quelli per i quali si ha il crollo nella distribuzione cumulativa dovuto al passaggio dalla coda 1TAT a quella 2TAT. Tali valori

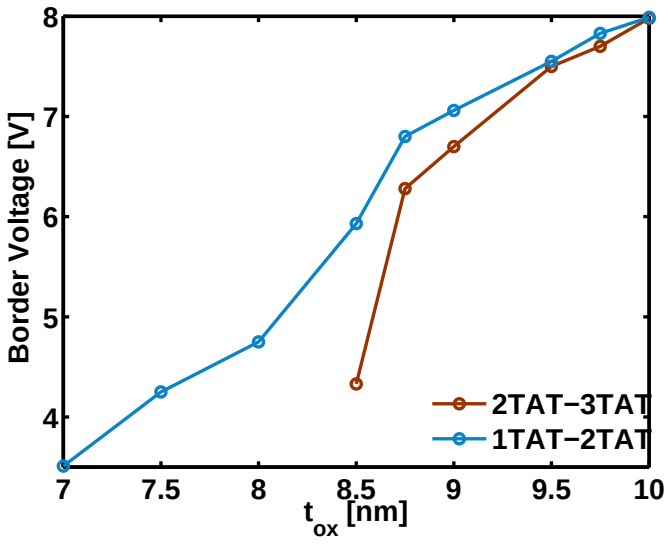


Figura 4.10: *Tensioni di "confine" che individuano il crollo delle distribuzioni cumulative dovuto al passaggio dal regime 1TAT al 2TAT e dal 2TAT al 3TAT, in funzione dello spessore del tunnel oxide.*

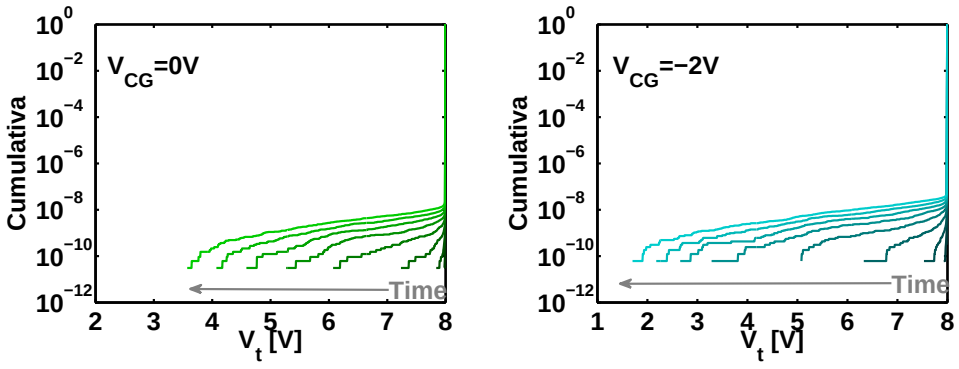
sono di approssimativamente $8.9nm$ per la tensione critica di $7V$, $9.45nm$ per $7.5V$ e $9.85nm$ per $7.9V$.

§4.3 Accelerazione in tensione

All'atto pratico, è decisamente scomodo dover effettivamente aspettare dieci anni per verificare la reale tenuta delle celle. Per questo si sono escogitati dei metodi di misura che consentissero di "accelerare" il tempo efficace esperito dalla cella, in modo da replicare l'evoluzione conseguita in dieci anni in un tempo di al massimo pochi giorni.

Questi metodi consistono nel polarizzare l'elettrodo di *control gate* non più ad una tensione nulla, ma ad una tensione sufficientemente negativa in grado di accelerare il processo di perdita di carica dal *floating gate* verso il substrato, portando la cella ad essere fallimentare in breve tempo. Campionando il tempo di fallimento a diverse tensioni si può così pensare di ricavare, tramite interpolazione, il lasso di tempo approssimato che deve realmente trascorrere nelle condizioni di ritenzione a $V_{CG} = 0V$.

Operativamente si procede come segue. Si fissano innanzi tutto dei criteri di fallimento: consideriamo al solito una *fail probability* di 10^{-9} alle tensioni critiche di $7V$, $7.5V$ e $7.9V$. Si fissa poi la densità di trappole, dal momento che sto considerando sempre lo stesso *array*, e si va cercare, al variare della tensione di *control gate* per quale tempo la distribuzione cumulativa interseca



(a) (b)

Figura 4.11: *Distribuzioni cumulative campionate a diversi tempi, equispaziati logaritimicamente tra 1 secondo e 10 anni. Entrambe fanno riferimento ad un* $t_{ox} = 7nm$ *e una* $N_T = 10^{10}cm^{-3}$*, ma mentre per (a)* $V_{CG} = 0V$*, per (b)* $V_{CG} = -2V$*.*

il punto di fallimento. Degli esempi di come queste distribuzioni cambino nel tempo sono mostrati nelle Figs. 4.11(a) e 4.11(b), che fanno riferimento alle condizioni di $t_{ox} = 7nm$, $N_T = 10^{10}cm^{-3}$, mentre $V_{CG} = 0V$ e $-2V$ rispettivamente. Come si vede, l'accelerazione in tensione è evidente: le curve per $V_{CG} = -2V$ sono molto più allungate rispetto alle corrispettive a $V_{CG} = 0V$, *i.e.* intersecano il punto di fallimento in un tempo minore. Da notare anche come le code delle curve "accelerate" riescano a raggiungere tensioni inferiori ai $3V$, soglia neutra originaria, e questo perché dal *floating gate* viene persa non solo tutta la carica immagazzinata nell'operazione di programmazione, ma anche dell'ulteriore carica propria del semiconduttore, arrivando ad avere nel FG una carica netta positiva anziché negativa. Rappresentando infine i tempi di fallimento in funzione della tensione di *control gate*, si ottengono i grafici mostrati nelle Figs. dalla 4.12(a) alla 4.12(d). Essi fanno riferimento ad una concentrazione $N_T = 10^{10}cm^{-3}$ per $t_{ox} = 7nm$ (a) e $8.5nm$ (b), mentre $N_T = 10^{11}cm^{-3}$ per $t_{ox} = 9nm$ (c) e $10nm$ (d). Il voltaggio relativo a ciscuna curva indica invece la tensione di fallimento prescelta. Come si può constatare, in un grafico semilogaritmico l'andamento del tempo di fallimento è sostanzialmente lineare con la tensione di CG. Questo dimostra che è sensato procedere al campionamento del suddetto tempo di *fail* per tensioni di *control gate* sufficientemente basse da garantire la fattibilità e poi interpolare linearmente per poter ricavare una stima alla tensione nulla, vero valore assunto dall'elettrodo in condizione di ritenzione.

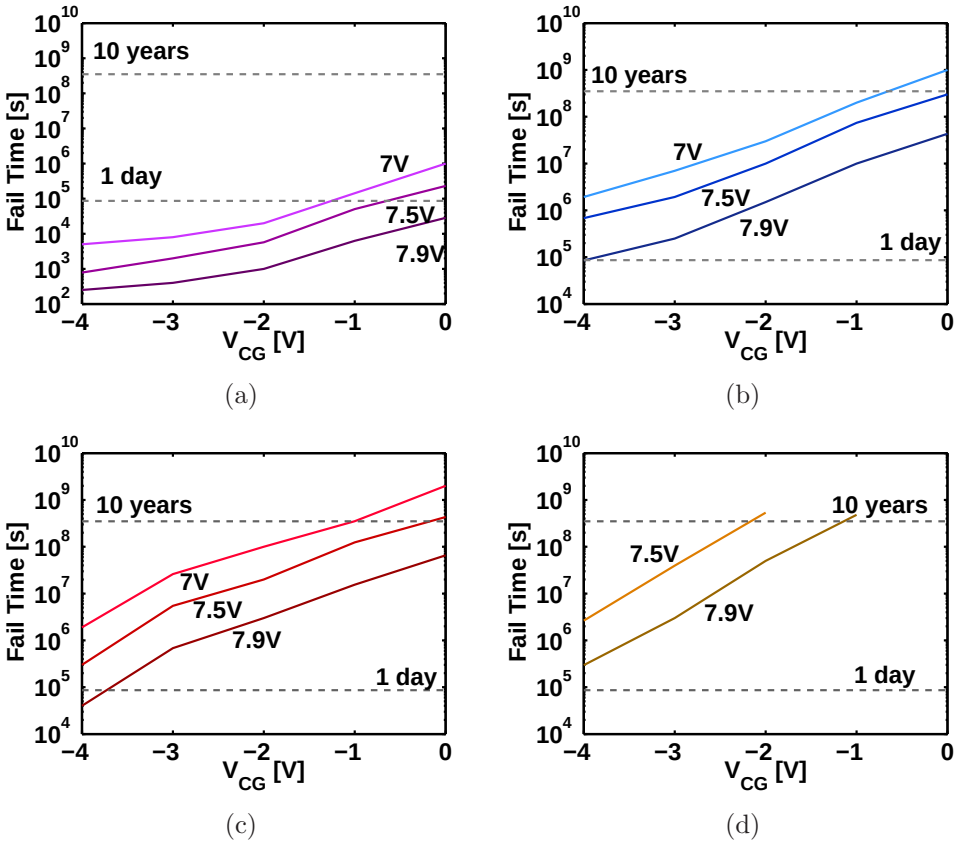


Figura 4.12: *Tempi di fallimento in condizione di accelerazione in funzione della tensione di control gate. In (a), $t_{ox} = 7nm$, e (b), $t_{ox} = 8.5nm$, abbiamo $N_T = 10^{10}cm^{-3}$, mentre in (c), $t_{ox} = 9nm$, e (d), $t_{ox} = 10nm$, abbiamo $N_T = 10^{11}cm^{-3}$. Le tensioni indicate a fianco di ogni curva fanno riferimento al criterio di fallimento prescelto.*

§4.4 Conclusioni

In questo capitolo abbiamo introdotto il concetto di memoria FLASH multilivello, individuandone i vantaggi ma anche i principali limiti e problemi. Per uno di questi, l'affidabilità in ritenzione, è stato condotto uno studio approfondito in merito alla dipendenza della densità di difetti critica nel *tunnel oxide*, da cui dipende l'entità della corrente TAT che "svuota" il *floating gate* precludendo la garanzia di non fallimento (con una certa probabilità) in un tempo standard di dieci anni, dallo spessore del *tunnel oxide* nel caso di architettura NOR. Ne è stato messo in luce l'andamento asintotico al diminuire dello spessore, mentre sono stati spiegati i repentini aumenti di concentrazione per determinati spessori come conseguenza del cambio di regime efficace di TAT. Inoltre, è stata considerata positivamente la tecnica dell'accelerazione in tensione come metodo per valutare il tempo di fallimento in condizioni di ritenzione standard a partire da dati relativi a tensioni di *control gate* più negative rispetto ad esse, individuando come sostanzialmente lineare la legge che lega il logaritmo del tempo di fallimento con la tensione V_{CG}.

Infine, occorre precisare un particolare. La determinazione della densità critica di difetti potrebbe non fornire informazioni sufficienti in merito all'affidabilità delle celle, in quanto essa non è un parametro facilmente conoscibile *a priori*. Bisogna cioè creare un modello che sia in grado di predire quale sia il suo valore a seguito di un certo tipo di utilizzo in termini di cicli di programmazione/cancellazione, che sono le operazioni di cui si ha il controllo. Tutto ciò, esulando ovviamente dagli scopi di questa tesi, è da intendersi solamente come suggerimento per ulteriori lavori di approfondimento.

Conclusioni Generali

In questo lavoro di tesi si sono introdotti i concetti base riguardo alle memorie non volatili a stato solido di tipo FLASH, indicandone le principali caratteristiche elettriche ed architetturali, con uno sguardo alle opportunità e alle difficoltà che il continuo *scaling* dei dispositivi comporta. Si sono anche messe in luce le possibilità dell'utilizzo di materiali ad elevata permettività dielettrica per limitare le correnti di *leakage* di *gate*, fornendo infine un elenco dei principali candidati a questo scopo. Si sono poi analizzati ed elaborati alcuni modelli per il calcolo della corrente di *tunneling* diretto/Fowler–Nordheim basato sull'approssimazione semi–classica WKB e per la corrente *Trap Assisted*, principale causa, quest'ultima, dei problemi di affidabilità per strutture MOS ultra scalate. I modelli TAT monodimensionali ad una e due trappole sono stati utilizzati per determinare i parametri di *fitting* di misure sperimentali relative a caratteristiche IV di un MOS-HK al variare della temperatura. Il modello TAT è stato poi ampliato ad un numero qualsiasi di difetti e ad un ambiente tridimensionale, passando da una visione delle trappole sottoforma di densità ad una che invece le concepisce puntiformi. Sono state analizzate le approssimazioni che hanno reso possibile questa evoluzione, verificando poi con successo la coerenza tra i modelli monodimensionali e tridimensionali in merito al calcolo delle correnti di perdita. Questo modello di TAT è stato successivamente impiegato in un contesto di simulazioni Monte Carlo, dopo aver formalizzato le tecniche di generazione di numeri *random* e di *enhancement* statistico utilizzate, al fine di valutare la statistica delle correnti di *gate* in un *array* di memorie FLASH al variare dello spessore del *tunnel oxide*. Anche in questo caso, il *fitting* dei dati sperimentali ha permesso di valutare i parametri dei difetti responsabili della perdita, ovvero le distribuzioni di probabilità nello spazio e nell'energia e il tipo di correlazione esistente per quanto riguarda la loro generazione. Infine, sempre sfruttando gli strumenti simulativi sviluppati, è stato condotto uno studio sulla statistica della tensione di soglia di un banco di memorie FLASH NOR multilivello in ritenzione, analizzando in particolare la dipendenza della densità di difetti critica, secondo diversi criteri di fallimento, dallo spessore dell'ossido di *tunnel*, variabile dagli attuali $10nm$ a spessori inferiori. È stato anche valutato il metodo dell'accelerazione di tensione come valida tecnica per stimare il tempo di fallimento in condizione di ritenzione

standard di un *array* mediante interpolazione dei tempi di fallimento in condizione di perdita, per l'appunto, accelerata.
I modelli e gli strumenti di calcolo numerico sviluppati in questo lavoro sono, tuttavia, flessibili e possono essere impiegati per ulteriori lavori di approfondimento sia sui meccanismi che governano la corrente di *leakage*, in dispositivi MOS–HK e in celle di memoria FLASH, sia sulla statistica di altre grandezze di interesse per quanto riguarda l'affidabilità di banchi di memorie.

Bibliografia

[1] L. Larcher, P. Pavan, L. Albani and T. Ghilardi, "Bias and W/L Dependence of Capacitive Coupling Coefficients in Floating Gate Memory Cells", *IEEE Transaction on Electron Devices*, vol. 48, no. 9, pp. 2081–2089, September 2001.

[2] P. Cappelletti, C. Golla, P. Olivo and E. Zanoni, *FLASH MEMORIES*, Kluwer Academic Publishers, 1999.

[3] D. Ielmini, A. Spinelli, A. L. Lacaita, L. Confalonieri, and A. Visconti, "New tecnique for fast caracterization of SILC distribution in FLASH arrays", *ieee Reliability Physics Symposium*, pp. 73–80, 2001.

[4] B. Govoreanu, P. Blomme, M. Rosmeulen, J. Van Houdt, and K. D. Meyer, "VARIOT: a novel multilayer tunnel barrier concept for low–voltage nonvolatile memory devices", *IEEE Electron Device Letters*, vol. 24, no. 2, pp. 99–101, 2003.

[5] O. Engstrom, B. Raeissi, S. Hall, O. Buiu, M.C. Lemme, H.D.B. Gottlob, P.K. Hurley, K. Cherkaoui, "Navigation aids in the search for future high–k dielectrics: Physical and electrical trends",*Solid State Electronics*, vol. 51, pp. 622–626, 2007.

[6] G. D. Wilk, R. M. Wallace, J. M. Anthony,"High–κ gate dielectrics: Current status anc material properties considerations", *Journal of Applied Physics*, vol. 89, no. 10, pp. 5243–5275, 2001.

[7] J. Kwo, M. Hong, A. R. Kortan, K. L. Queeney, Y. J. Chabal, R. L. Opila, Jr., D. A. Muller, S. N. G. Chu, B. J Sapjeta, T. S. Lay, J. P. Mannaerts, T. Boone, H. W. Krautter, J. J. Krajewski, A. M. Sergent, and J. M. Rosamilla,"Properties of higk κ gate dielectrics Gd_2O_3 and Y_2O_3 for Si", *Journal of Applied Physics*, vol. 89, no. 7, pp. 3920–3927, 2001.

[8] P. W. Peacocka and J. Robertson, "Band offsets and Schottky barrier heights of high dielectric constant oxides", *Journal of Applied Physics*, vol. 92, no. 8, pp. 4712–4721.

[9] E. Rosencher and B. Winter, "Optoelectronics", Cambridge University Press, 2002.

[10] L. D Landau and E. M Lifsits, "Meccanica Quantistica. Teoria Non Relativistica", Editori Riuniti, 1975.

[11] M. Städele, B. R. Tuttle, and K. Hess, "Tunneling through ultrathin SiO_2 gate oxides from microscopic models", *IEEE Transaction on Electron Devices*, no. 44, pp. 1169–1171, 1997.

[12] R. Moazzani, and C. Hu, "Stress–induced leakage current in thin silicon dioxide films", *IEDM Technical Digest*, pp. 139–142, 1992.

[13] B. Riccò, G. Gozzi, and M. Lanzoni, "Modeling and simulation of stress–induced leakage current in ultrathin SiO_2 films", *IEEE Transaction on Electron Devices*, no. 7, pp. 1554–1560, 1998.

[14] D. Ielmini, A. Spinelli, M. Rigamonti, and A. Lacaita, "Modeling of SILC based on electron and hole tunneling – Part I", *IEEE Transaction on Electron Devices*, no. 6, pp. 1258–1264, 2000.

[15] S. Takagi, N. Yasuda, and A. Toriumi, "A new I–V model for Stress Induced Leakage Current including Inelastic Tunneling", *IEEE Transaction on Electron Devices*, vol. 68, no. 3, pp. 1059–1069, 1990.

[16] A. Palma, A. Godoy, J. A. Jiménez–Tejada, J. E. Carceller, and J. A. Lopez–Villanueva, "Quantum two–dimensional calculation of time constants of random telegraph signals in metal–oxide–semiconductor structures", *Physical Review B*, vol. 56, no. 15, pp. 9565–9574, 1997.

[17] D. Ielmini, A. S. Spinelli, M. A. Rigamonti, and A. L. Lacaita, "Modeling of SILC Based on Electron and Hole Tunneling–Part I: Transient Effects", *IEEE Transaction on Electron Devices*, vol. 47 , no. 6 , pp. 1258–1265, 2005.

[18] D. Ielmini, A. S. Spinelli, M. A. Rigamonti, and A. L. Lacaita, "Modeling of SILC Based on Electron and Hole Tunneling–Part II: Steady State", *IEEE Transaction on Electron Devices*, vol. 47 , no. 6 , pp. 1266–1272, 2005.

[19] F. Driussi, F. Widdershoven, D. Esseni, L. Selmi, and M. J. van Duuren. "Experimental Characterization of Statistically Independent Defects in Gate Dielectrics–Part I: Description and Validation of the Model", *IEEE Transaction on Electron Devices*, vol. 52, no. 5, pp. 942–948, 2005.

[20] F. Driussi, F. Widdershoven, D. Esseni, L. Selmi, and M. J. van Duuren. "Experimental Characterization of Statistically Independent Defects in Gate Dielectrics–Part II: Experimental Results on Flash Memory Arrays", *IEEE Transaction on Electron Devices*, vol. 52, no. 5, pp. 949–954, 2005

[21] W. Shockley, W. T. Read, Jr., "Statistics of the Recombination of Holes and Electrons", *Physical Review*, vol. 87, no. 5, pp. 835–842, 1952.

[22] R. Hall, "Electron–Hole Recombination in Germanium" *Physical Review*, vol. 87, p. 387, 1952.

[23] S. Takagi, N. Yasuda, and A. Toriumi, "Experimental Evidence of Inelastic Tunneling in Stress–Induced Leakage Current", *IEEE Transaction on Electron Devices*, vol. 46 , no. 2 , pp. 335–341, 1999.

[24] D. Goguenheim and M. Lannoo, "Theorethical and experimental aspects of the thermal dependence of electron capture coefficients", *Journal of Applied Physics*, vol. 68, no. 3, pp. 1059–1069, 1990.

[25] E. Driussi, D. Esseni, L. Selmi, M.J. van Duuren and E. Widdershoven, "Experimental Evidence and Statistical Modeling of Cooperating Defects in Stressed Oxides", Proc. of ESSDERC, p. 209–212, 2004.

[26] D. Ielmini, A. S. Spinelli, A. L. Lacaita, and Alberto Modelli, "A Statistical Model for SILC in Flash Memories", *IEEE Transaction on Electron Devices*, vol. 49, no. 11, 2002.

[27] J. Sune, E. Y. Wu, S. Tous, "A physics–based deconstruction of the percolation model of oxide breakdown", *Microelectronic Engineering*, 84, pp. 1917–1920, 2007.

[28] B. S. Shim, S. Jin, Y. J. Park, and H. S. Min, "A New Statistical Model for SILC Distribution of Flash Memory and the Effect of Spatial Trap Distribution", SISPAD 2006.

[29] D. Ielmini, A. S. Spinelli, A. L. Lacaita, A. Modelli, "Modeling of anomalous SILC in flash memories based on tunneling at multiple defects", *Solid State Electronics*, 46, pp. 1749–1756, 2002.

[30] D. Ielmini, A. S. Spinelli, A. L. Lacaita, A. Modelli, "A new two–trap tunneling model for the anomalous stress–induced leakage current (SILC) in Flash memories", *Microelectronic Engineering*, 59, pp. 189–195, 2001.

[31] R. Entner, A. Gehringt, H. Kosinat, T. Grasser, and S. Selberherrt, "Modeling of Tunneling Currents for Highly Degraded CMOS Devices"

[32] J. Westlinder, G. Sjoblom, J. Olsson, "Variable work function in MOS capacitors utilizing nitrogen–controlled TiNx gate electrodes", *Microelectronic Engineering*, 75, pp. 389–396, 2004.

[33] S. Monaghan, P.K. Hurley, K. Cherkaoui, M.A. Negara, A. Schenk, "Determination of electron effective mass and electron affinity in HfO_2 using MOS and MOSFET structures", *Solid State Electronics*, 53, pp. 438–444, 2009.

[34] A. Pacelli, " Self consistent solution of the Schrödinger equation in quantum wells by implicit iteration", *IEEE Transaction on Electron Devices*, no. 44, pp. 1169–1171, 1997.

[35] S. Fleischer, P. T. Lai, Y. C. Cheng, "Simplified closed–form trap–assisted tunneling model applied to nitrided oxide dielectric capacitors", *Journal of Applied Physics*, vol. 72, no. 12, pp. 5711–5715.

[36] L. Devroye, "Non–Uniform Random Variate Generation", New York: Springer–Verlag. Cap. 2, Sez. 2, p. 28, 1986.

[37] G. E. P. Box and M. E. Muller, "A Note on the Generation of Random Normal Deviates", *The Annals of Mathematical Statistics*, Vol. 29, No. 2, pp. 610–611, 1958.

[38] D. E. Knuth, "Seminumerical Algorithms", Addison Wesley, 1969.

[39] G. Rubino, B. Tuffin, "Rare Event Simulation using Monte Carlo Methods", John Wiley & Sons, Ltd, 2009.

[40] D. Ielmini, A. S. Spinelli, A. L. Lacaita and A. Modelli, "A Statistical Model for SILC in Flash Memories", *IEEE Transaction on Electron Devices*, vol. 49, no. 11, pp. 1955–1961, 2002.

[41] B. Eitan, R. Kazerounian, A. Roy, G. Crisenza, P. Cappelletti and A. Modelli, "Multilevel Flash cells and their Trade–offs", *IEDM Technical Digest*, pp. 169–172, 1996.