Università degli Studi di Padova

Dipartimento di Ingegneria

Corso di Laurea in Ingegneria dell'Informazione

Memorie Flash a NAND

Laureando **Henry Felber**

Relatore **Prof. Andrea Gerosa**

Anno Accademico 2013/2014

A mio padre, senza il quale non avrei questa passione, mia madre, per il suo continuo sostegno, mia sorella e Carlo.
Ai miei amici, per sopportarmi ogni giorno.
CIATT.

> A process cannot be understood by stopping it. Understanding must move with the flow of the process, must join it and flow with it.
>
> *Frank Herbert,* Dune

Indice

Sommario

In questa tesi verrà analizzata una delle tecnologie più diffuse, e allo stesso tempo più efficaci, dell'era moderna: le memorie flash.
Partendo da una breve introduzione all'argomento, si studierà prima il funzionamento a livello fisico del transistor, il suo inserimento in un'architettura a NAND e i metodi di accesso e scrittura del dati in tale contesto. Verranno poi discussi l'affidabilità ed i principali problemi della tecnologia, insieme alle possibili soluzioni ad essi. Per concludere, una panoramica delle principali applicazioni delle memorie Flash nei nostri giorni, con un particolare sguardo al futuro della tecnologia e agli sviluppi possibili.

Capitolo 1

Introduzione

Negli ultimi anni abbiamo assistito, soprattutto grazie alla diffusione di dispositivi quali telefoni cellulari, lettori mp3, macchine fotografiche digitali, penne USB, e altri apparecchi elettronici di vario tipo, ad un fenomeno esplosivo e a dir poco virale nel campo delle memorie a semiconduttore: la crescita delle memorie Flash.
In futuro avremo sempre più bisogno di un maggior numero di memorie non volatili, con alta densità e alte velocità di scrittura per applicazioni nel campo della conservazione di dati, o accessi casuali veloci nel caso di esecuzione di codice in loco.
Le prime memorie Flash risalgono circa agli inizi degli anni '90, e ciononostante essa è una tecnologia ancora estremamente sviluppabile in vari campi, soprattutto grazie all'estrema conoscenza in materia, alla flessibilità e ai costi contenuti rispetto alle simili EPROM ed EEPROM: i dati di vendita nel campo delle memorie a semiconduttore dimostrano l'affidabilità, il largo utilizzo e la maturità di questa tecnologia nella maggior parte delle applicazioni delle memorie non volatili [Figura 1.1].

Le memorie a semiconduttore possono essere divise, fondamentalmente, in due rami principali:

- Le memorie volatili, come ad esempio le SRAM o DRAM, che, nonostante la grande velocità di scrittura e lettura (SRAM) o la densità (DRAM), perdono il dato immagazzinato quando l'alimentazione è scollegata.

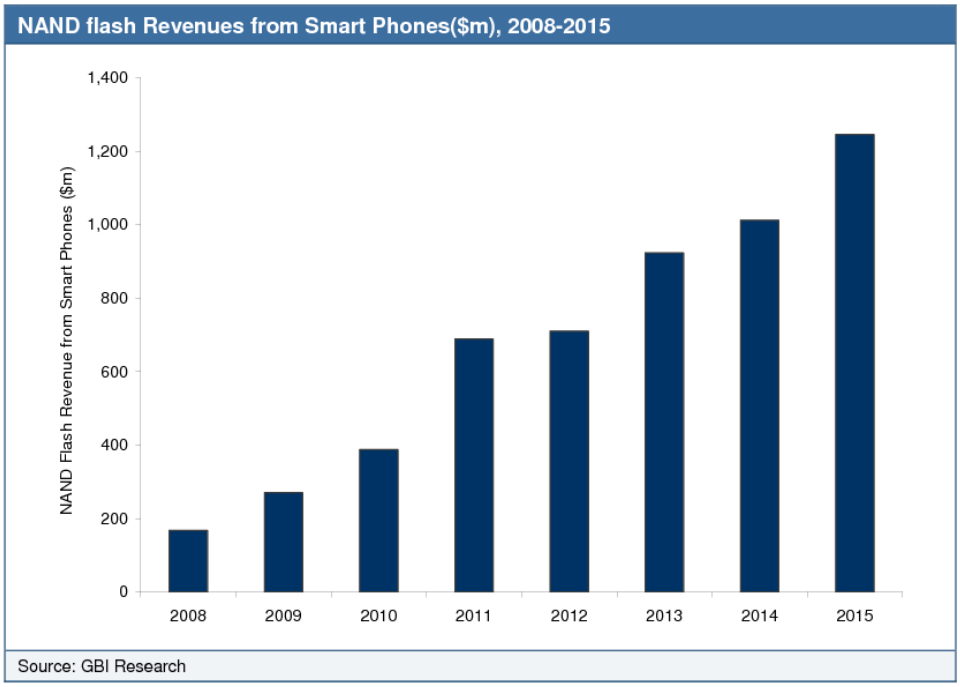

Figura 1.1: Dati di vendita delle memorie Flash a NAND per smartphone e tablet

- Le memorie non volatili, come EPROM, EEPROM, o appunto le Flash, che bilanciano le minori prestazioni in lettura e scrittura con la capacità di mantenere i dati anche in assenza di alimentazione.

E' quindi grazie a questa capacità di non perdere il dato scritto che le memorie non volatili hanno offerto varie opportunità e coprono un largo campo di applicazioni.
Le memorie non volatili possono essere comparate qualitativamente secondo flessibilità e costo [Figura 1.2].

Per flessibilità intendiamo la possibilità di essere programmate, cancellate e riscritte su sistema a diversi livelli di precisione (il chip intero, una singola pagina, un byte, un bit).
Per costo invece parliamo della complessità del processo di produzione e in particolare la densità del silicio, ovvero la dimensione delle singole celle. Dalla Figura 1.2 notiamo come le memorie Flash offrano le migliori prestazioni come compromesso tra i due parametri, dal momento che hanno la minore dimensione di cella (1 Transistor, al contrario dei 2 Transistor per le EE-

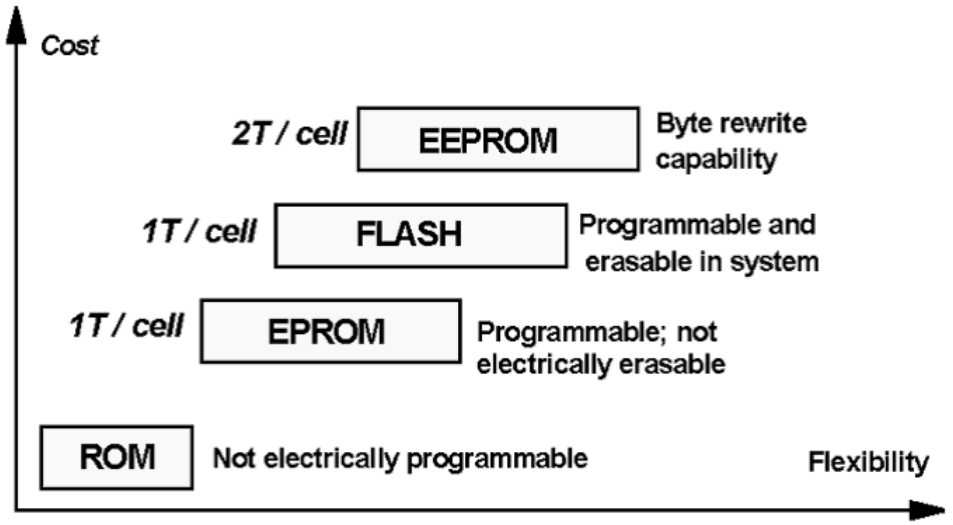

Figura 1.2: Comparazione delle memorie non volatili secondo costo e flessibilità

PROM) e un'ottima flessibilità (possono essere riscritte fino a 10^5 volte, con la programmazione per byte e la cancellazione per settori).
Anche nel campo delle memorie Flash possiamo trovare un'ulteriore suddivisione, a seconda dell'architettura di funzionamento e, di conseguenza, del principale campo di utilizzo:

- Flash a NOR, che consentono di scrivere e e leggere singoli byte, e dispongono di tempi di accesso e scrittura relativamente brevi: vengono quindi principalmente usate, per la semplicità di programmazione, come sostitute delle EPROM, nell'esecuzione di codice già precedentemente scritto.

- Flash a NAND, che consentono la lettura a blocchi: possono quindi essere indirizzati solo blocchi di dati, chiamati pagine. Il vantaggio principale di questa architettura è l'incredibile densità, inferiore di 2.5 volte a quella della NOR, essendo ogni cella collegata alla stessa Bit Line e occupando quindi minore spazio. Le Flash a NAND trovano maggior utilizzo nel campo di immagazzinamento dati, avendo una struttura simile a quella degli Hard Disk.

E' grazie allo sviluppo di apparecchi portatili che le memorie NAND hanno preso sempre più piede, surclassando le NOR: nel 2003 vi è stato il sorpasso

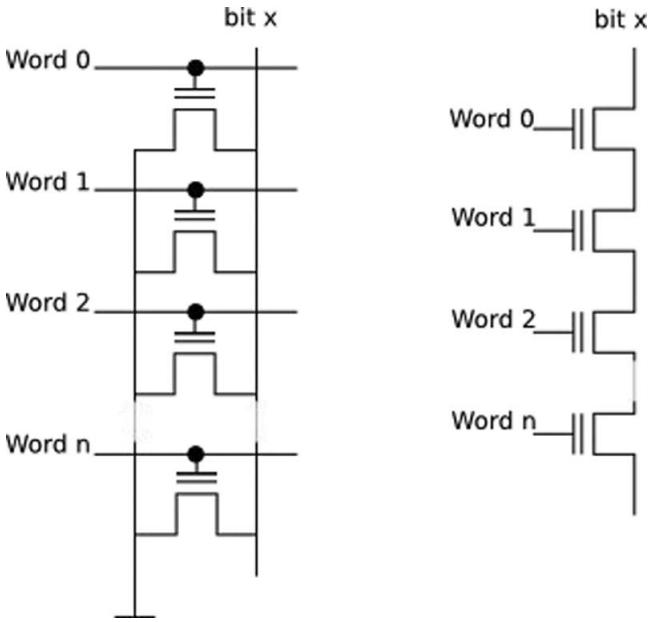

Figura 1.3: Architetture NOR (sinistra) e NAND (destra) a confronto

in termini di vendite, nel 2005 le memorie a NAND hanno superato anche le DRAM, e nel solo 2007 sono stati acquistati più byte di memorie Flash di quanti byte siano mai stati acquistati, in totale, di DRAM.

In questa tesi mi concentrerò principalmente su quest'ultima applicazione delle memorie Flash, l'architettura a NAND.
Nel capitolo 2 parlerò del funzionamento della singola cella, delle modalità di accesso e scrittura e mi focalizzerò sull'architettura a NAND.
Nel capitolo 3 affronterò le principali problematiche della tecnologia, e i possibili metodi di risoluzione per renderle ancora più affidabili.
Nel capitolo 4 parlerò delle maggiori applicazioni delle memorie Flash a NAND, con un occhio di riguardo per le ormai diffuse SSD e per le schede di memoria.
Nel capitolo 5 trarrò le conclusioni finali con uno sguardo al futuro di questo tecnologia e sui possibili miglioramenti.

Capitolo 2

Funzionamento delle memorie Flash

2.1 Il transistor a Gate Flottante

Vi sono diversi tipi di celle per una memoria, e ognuna si differenza per una caratteristica principale, sia essa l'energia assorbita, il tempo e la velocità di programmazione, la selettività, e così via: per una memoria Flash vi era la necessità di memorizzare informazioni in grado di variare nel corso del tempo, potendo essere cancellate e riscritte, e allo stesso tempo si necessitava di una struttura in grado di garantire la durabilità, ovvero che il dato, anche dopo lunghi periodi, non venisse perso ma anzi fosse immagazzinato in maniera efficace.
Per poter immagazzinare informazioni indipendentemente dalle condizioni esterne, e poter commutare da uno stato all'altro, abbiamo bisogno di una struttura la cui conduttività possa variare. La miglior soluzione è un transistor con una V_T, una tensione di soglia, che possa variare, a seconda delle esigenze, tra due diversi stati, corrispondenti ad esempio ai valori logici "0" e "1" del bit da salvare: le celle possono quindi avere questi due diversi valori o programmandole o cancellandone il contenuto.

Esistono diversi metodi per l'immagazzinamento dei dati:

- Viene cancellata l'intera memoria, e viene programmato il singolo bit selezionato.
- Viene cancellato e riprogrammato con le nuove informazioni solo il byte corrispondente al bit che vogliamo cambiare.
- Viene indirizzato solo il bit da cambiare: se contiene già l'informazione che vogliamo immagazzinare, essa viene mantenuta, altrimenti viene sovrascritta.

Nei primi due casi c'è un'unica operazione attuabile bit per bit, ed è chiamata "programmazione"; l'operazione attuabile su tutto l'array è invece chiamata "cancellazione". Nel terzo caso invece entrambe le operazioni possono essere effettuate bit per bit ma la programmazione necessita di un'organizzazione più complicata dell'array. L'operazione base rimane sempre e comunque la variazione della tensione di soglia della cella di memoria, che viene effettuata cambiando il numero di cariche presenti tra il Gate ed il canale: esprimiamo V_T come

$$V_T = K - \frac{\bar{Q}}{C_{ox}} \tag{2.1}$$

dove K è una costante che dipende dai materiali del Gate e del substrato, dal drogaggio e dallo spessore dell'ossido di Gate, mentre $\bar{Q}$ è la carica pesata e C_{ox} è la capacità dell'ossido di Gate.
Agiamo sul fattore $\frac{\bar{Q}}{C_{ox}}$, cambiando cioè il numero di cariche presenti, per influenzare la tensione di soglia: l'immagazzinamento delle cariche può avvenire in vari metodi, a seconda del dispositivo. In questa tesi verrà trattato il Floating Gate Transistor (FGMOS), dove le cariche sono concentrate in uno strato di materiale conduttivo tra Gate e canale, completamente circondato da isolante.
Un modo per comparare due diverse celle è la loro efficienza, descritta tramite la resistenza (capacità di mantenere il dato dopo vari cicli di lettura/scrittura) e la conservazione (capacità di mantenere il dato a lungo nel tempo).
I MOSFET (Metal-Oxide-Semiconductor Field Effect Transistor) non vengono più usati proprio per la scarsa affidabilità nei due campi precedentemente descitti: gli FGMOS invece sono la base di ogni moderna memoria non volatile, in particolare delle Flash.
Il Floating Gate Transistor, dunque, è un elemento dalla struttura simile al MOSFET, in cui abbiamo un Gate, completamente circondato da dielettrico, chiamato appunto Floating Gate, il classico Control Gate, che governa il

Gate flottante variando di tensione, e il substrato.

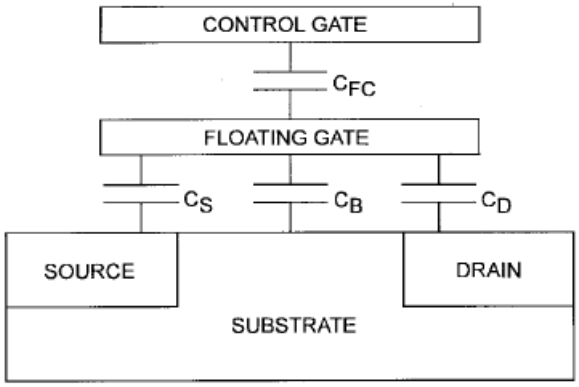

Figura 2.1: Struttura di un FGMOS

Storicamente, gli inventori del FGMOS sono ritenuti Dawon Kahng e Simon Sze, che scoprirono questa tecnologia nel 1967: una struttura in grado di immagazzinare carica in modo semipermanente.
Il Floating Gate è elettricamente isolato per permettergli di diventare l'elettrodo di carica della cella: l'apparente variazione della V_T del transistor della cella è dovuta alla carica iniettata nel Floating Gate e ivi mantenuta dal dielettrico, la cui qualità garantisce la non volatilità del dato, mentre lo spessore del dielettrico stesso definisce la possibilità di programmare o cancellare la cella tramite impulsi elettrici.
L'isolamento del Floating Gate dal substrato è attuato con uno strato isolante di SiO_2, chiamato "ossido di tunnel" poichè l'effetto tunnel di Fowler-Nordheim avviene attraverso esso, abbastanza sottile (circa 9-10 nm) da permettere l'iniezione e il passaggio di cariche verso il Floating Gate, mentre la separazione dal Control Gate avviene tramite un triplo strato di ONO (Oxide - Nitride - Oxide), sufficientemente spesso (15-20 nm) da non permettere la perdita di carica dal Floating Gate al Control Gate e aumentare la qualità dell'ossido di tunnel.
In Figura 2.1 è rappresentata la struttura di un FGMOS: C_{FC}, C_S, C_D e C_B sono le capacità presenti tra Control Gate e Floating Gate, e tra Floating Gate e, rispettivamente, Source, Drain e Bulk.
Partiamo dalla descrizione di un FGMOS inizialmente scarico, ovvero dove $\bar{Q} = 0$: abbiamo

$$\bar{Q} = 0 = C_{FC}(V_{FG} - V_{CG}) + C_S(V_{FG} - V_S) + C_D(V_{FG} - V_D) + C_B(V_{FG} - V_B) \quad (2.2)$$

dove V_{FG} è il potenziale sul Floating Gate, V_{CG} il potenziale sul Control Gate, V_S sul Source, V_D sul Drain e V_B sul Bulk.
Definiamo $C_T = C_{FC} + C_S + C_D + C_B$ come la capacità totale del Floating Gate, e definiamo invece $\alpha_J = C_J/C_T$ il coefficiente di accoppiamento relativo all'elettrodo J, dove J può essere uno a scelta tra Gate, Source, Drain o Bulk: possiamo quindi esprimere il potenziale del Floating Gate nella seguente maniera:

$$V_{FG} = \alpha_G V_{GS} + \alpha_D V_{DS} + \alpha_S V_S + \alpha_B V_B \quad (2.3)$$

Notiamo che nella (2.3) il potenziale del Floating Gate non dipende solo da quello del Control Gate, ma anche dagli altri elettrodi: supponendo Bulk e Source a massa, possiamo riscrivere la (2.3) in

$$V_{FG} = \alpha_G(V_{GS} + \frac{\alpha_D}{\alpha_G} V_{DS}) = \alpha_G(V_{GS} + f \cdot V_{DS}) \quad (2.4)$$

dove abbiamo posto

$$f = \frac{\alpha_D}{\alpha_G} \quad (2.5)$$

Le equazioni per il Floating Gate Transistor si possono ottenere da quelle del normale MOSFET sostituendo al potenziale di Gate V_{GS} quello del Floating Gate V_{FG} e sostituendo i parametri di dispositivi come la tensione di soglia V_T e il fattore di conduttività β con i valori corrispettivi del FGMOS misurati rispetto al Control Gate.
Definendo, per $V_{DS} = 0$ (V_T^{FG} è la tensione di soglia del Floating Gate, V_T^{CG} quella del Control Gate)

$$V_T^{FG} = \alpha_G V_T^{CG} \quad (2.6)$$

e (β^{FG} è il fattore di conduttività del Floating Gate, β^{CG} del Control Gate)

$$\beta^{FG} = \frac{1}{\alpha}\beta^{CG} \quad (2.7)$$

possiamo comparare le due equazioni delle caratteristiche I-V per il classico MOSFET e per il FGMOS nella regione lineare e in quella di saturazione classica.
MOSFET:
Lineare

$$|V_{DS}| < |V_{GS} - V_T| \tag{2.8}$$

$$I_{DS} = \beta[(V_{GS} - V_T)V_{DS} - \frac{1}{2}V_{DS}^2] \tag{2.9}$$

Saturazione

$$|V_{DS}| \geq |V_{GS} - V_T| \tag{2.10}$$

$$I_{DS} = \frac{\beta}{2}(V_{GS} - V_T)^2 \tag{2.11}$$

FGMOS:
Lineare

$$|V_{DS}| < \alpha_G|V_{GS} + fV_{DS} - V_T| \tag{2.12}$$

$$I_{DS} = \beta[(V_{GS} - V_T)V_{DS} - (f - \frac{1}{2\alpha_G})V_{DS}^2] \tag{2.13}$$

Saturazione

$$|V_{DS}| \geq \alpha_G|V_{GS} + fV_{DS} - V_T| \tag{2.14}$$

$$I_{DS} = \frac{\beta}{2}\alpha_G(V_{GS} + fV_{DS} - V_T)^2 \tag{2.15}$$

Dove β e V_T nelle (2.12), (2.13), (2.14) e (2.15) sono riferite al Control Gate. Possiamo notare vari effetti in queste equazioni, dovuti in particolare all'accoppiamento capacitivo tra Drain e Floating Gate, principale differenza e

modificatore della caratteristica I-V tra le due diverse strutture dei transistor:

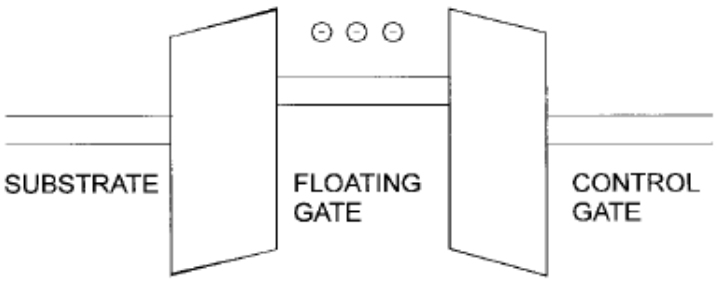

Figura 2.2: Diagramma del potenziale di un FGMOS

1. Il FGMOS può condurre corrente anche nel caso in cui si abbia $|V_{GS}| < |V_T|$. Questo succede poichè il canale può essere formato dal voltaggio al Drain attraverso il termine $f \cdot V_{DS}$ nella (2.12), effetto comunemente chiamato "drain turn-on".

2. La regione di saturazione in un MOSFET è quella in cui la I_{DS} è sostanzialmente indipendente dalla tensione applicata al Drain: ciò nel FGMOS non è più valido, la corrente aumenta all'incrementare del voltaggio di Drain e non satura.

3. La transconduttanza nella regione di saturazione è

$$g_m = \frac{\partial I_{DS}}{\partial V_{GS}}_{V_{DS}=constant} = \alpha_G \beta (V_{GS} + f V_{DS} - V_T) \tag{2.16}$$

   dove g_m aumenta con V_{DS} nel FGMOS al contrario del normale MOSFET, dove g_m è relativamente indipendente dal voltaggio di Drain nella regione di saturazione.

4. Il rapporto di accoppiamento capacitivo f dipende unicamente da C_D e C_{FG} ($f = \alpha_D/\alpha_G = C_D/C_{FC}$), e si può verificare il suo valore, nella regione di saturazione, con

$$f = -\frac{\partial V_{GS}}{\partial V_{DS}}_{I_{DS}=constant} \tag{2.17}$$

Per programmare e cancellare il dato nel FGMOS vengono usate varie soluzioni: il problema principale è trasferire la carica elettrica da e su il Floating Gate, forzando la carica a passare attraverso lo strato di materiale isolante (In Figura 2.2 notiamo come i dielettrici formino delle barriere di potenziale mano a mano che la carica viene immagazzinata nel Floating Gate).
Due sono i principali metodi utilizzati:

- Hot-Electron Injection (HEI): un campo laterale (tra Source e Drain) "eccita" gli elettroni e un campo elettrico trasversale (tra Control Gate e Bulk) inietta le cariche nel Floating Gate.

- Fowler-Nordheim Tunneling: meccanismo che inizia quando c'è un forte campo elettrico attraverso il sottile strato di ossido. In queste condizioni, il diagramma del potenziale dell'ossido è molto ripido: c'è quindi un'alta probabilità che gli elettroni passino attraverso la barriera di potenziale.

2.1.1 Hot-Electron Injection

Il funzionamento fisico del HEI è relativamente semplice: un elettrone viaggia dal Source al Drain, guadagnando energia dal campo elettrico laterale e perdendo energia per colpa delle vibrazioni del reticolo cristallino.
Con campi bassi, questa è una condizione di equilibrio dinamico, che viene mantenuta finchè il campo raggiunge circa i 100 kV/cm. Per i campi che superano questo valore, gli elettroni non sono più in equilibrio col reticolo e la loro energia inizia ad incrementare. Gli elettroni sono "eccitati" dal forte campo laterale, e una piccola porzione di loro ha abbastanza energia da superare la barriera tra l'ossido e il silicio: questo può avvenire a patto che vengano rispettate tre condizioni. Prima di tutto, devono possedere un'energia cinetica maggiore del potenziale della barriera; secondariamente, la loro direzione deve essere perpendicolare alla barriera; infine, devono essere raccolti dal campo nell'ossido.
Per sapere quanti elettroni riusciranno a superare la barriera di potenziale, dobbiamo conoscere la distribuzione di energia $f_E(\varepsilon, x, y)$ come funzione del campo laterale ε, la distribuzione della quantità di moto $f_k(E, x, y)$ in funzione dell'energia E dell'elettrone (ovvero quanti elettroni sono diretti verso l'ossido), la forma e l'altezza della barriera di potenziale, e la probabilità che un elettrone di energia E, vettore d'onda k, e distanza d dall'interfaccia Si/SiO_2 superi la barriera di potenziale. Ognuna di queste funzioni deve essere esplicitata in ogni punto del canale, anche se ciò richiederebbe un modello decisamente difficile da gestire: inoltre, quando l'energia guadagnata dagli elettroni raggiunge una determinata soglia, l'impatto col reticolo cristallino e con gli altri portatori influisce in maniera pesante sulla perdita di energia, rendendo obbligatoria la sua inclusione nelle nostre considerazioni. Possiamo comunque semplificare il nostro studio con due diversi modelli.
La corrente del HEI è spesso descritta e simulata seguendo il modello de "l'elettrone fortunato", non estremamente preciso ma che si è comunque rivelato affidabile e soprattutto semplice, basato sulla supposizione che un elettrone abbia abbastanza fortuna da viaggiare con un moto balistico nel campo ε per una distanza sufficientemente lunga, eventualmente acquisendo abbastanza energia da superare la barriera di potenziale se una collisione lo spinge in direzione dell'interfaccia Si/SiO_2.
Di conseguenza influiscono varie probabilità in questo calcolo, schematizzate nella Figura 2.3: la probabilità che il portatore sia abbastanza fortunato da superare la barriera di ossido, avendo acquisito energia dal campo laterale, e conservandola dopo la collisione (col reticolo o con un altro portatore) che gli ha consentito di acquisire la direzione corretta (P_{ϕ_b}); la probabilità che, lungo il percorso dal punto di collisione all'interfaccia, esso non subisca al-

tri urti, che potrebbero fargli perdere energia o modificare la sua direzione (P_{ED}); la probabilità che il portatore abbia energia a sufficienza per superare il campo dell'ossido, repulsivo, e che quindi tende a impedirgli il passaggio della barriera (P_{OC}).

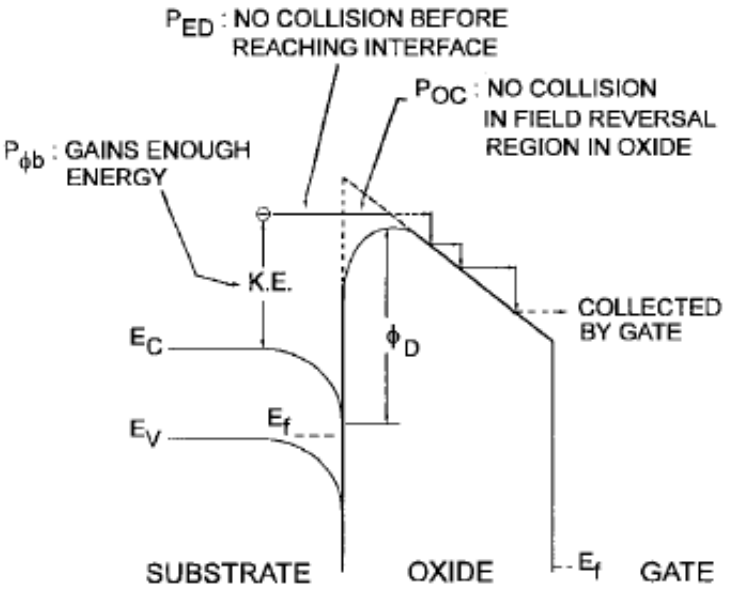

Figura 2.3: Un diagramma dell'energia con i tre processi nell'iniezione

L'altro modello invece assimila l'elettrone ad un gas in equilibrio termico, e lo caratterizza con una temperatura diversa da quella del reticolo, mettendola in relazione col campo di deriva: esso è un approccio sicuramente più formale, e mette in dipendenza la probabilità di superare la barriera con il profilo completo del campo elettrico in quella regione di canale, essendo una relazione non-locale.
Entrambi i modelli, comunque, consentono di fornire la seguente relazione tra corrente di substrato I_{sub} e la corrente di iniezione I_G

$$I_G/I_{ch} \sim I_{sub}/I_{ch} e^{-\Phi/\Phi_i} \tag{2.18}$$

con I_{ch} corrente di canale, Φ_i energia di ionizzazione da impatto, e Φ è l'energia della barriera vista dai portatori che devono essere iniettati nell'ossido: quest'ultima barriera deve essere corretta per effetto delle componenti di tunnel nella corrente di Gate.

La corrente di substrato è formata da lacune generate dall'impatto degli elettroni con il reticolo cristallino nella regione di Drain: esse sono sempre generate dal momento che la soglia d'energia di ionizzazione Φ_i ($\sim$ 1.6 V) è più bassa della barriera di energia di iniezione Φ ($\sim$ 3.2 V). Alcune lacune possono ottenere energia dal campo elettrico laterale per essere iniettate nell'ossido, facendolo degradare. Il processo di ionizzazione genera anche un gran numero di portatori che possono essere iniettati nelle regioni dell'ossido, dove possono essere intrappolati vicino all'interfaccia degradando le prestazioni del dispositivo.

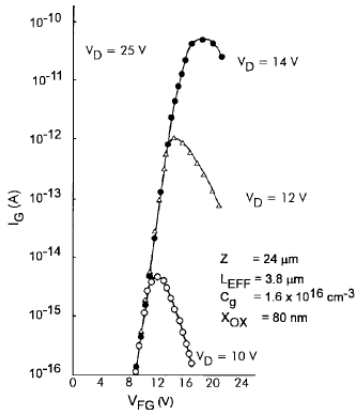

Figura 2.4: I_G in funzione del V_{FG} con V_D come parametro

La Figura 2.4 mostra la corrente d'iniezione I_G misurata in un FGMOS con una tecnica indiretta basata sulla relazione $I_G = dQ_{FG}/dt \approx C_{FG}\Delta V_{FG}\ /\Delta t$ dove Q_{FG} è la carica nel Floating Gate, C_{FG} è la capacità di accoppiamento tra il Floating ed il Control Gate, e V_{FG} è il potenziale del Floating Gate.
La forma di queste curve mostra la dipendenza del meccanismo di iniezione da V_G e V_D. Quando $V_G < V_D$, c'è un punto nel canale dove il campo elettrico trasversale nell'ossido cambia la sua direzione e respinge gli elettroni invece di attrarli: il punto di pinch-off è più vicino al Source rispetto a questo punto di inversione. All'incrementare di V_G il campo laterale medio aumenta, ma il punto di iniezione si sposta verso il Drain a causa del variare del punto di inversione, anche lui dipendente da V_G. Come risultato, gran parte degli elettroni disponibili per l'iniezione si trova in una regione dove il campo laterale è maggiore, portando ad un rapido decremento della I_G. Quando la V_G supera la V_D, la distribuzione degli elettroni eccitati è sottomessa al campo elettrico laterale nei pressi del Drain, che diminuisce all'aumentare di V_G, riducendo così la corrente di Drain. I_G dipende anche

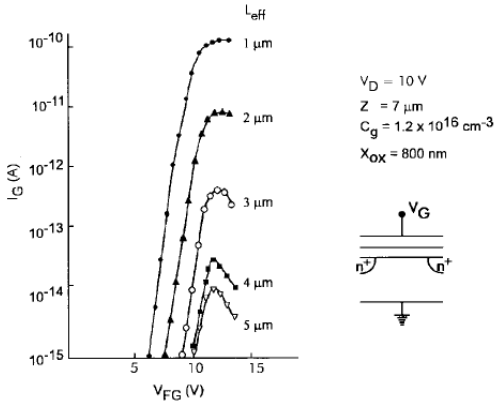

Figura 2.5: I_G in funzione del V_{FG} con L come parametro

dalla lunghezza di canale, come mostrato in Figura 2.5: una diminuzione di esso porta ad un aumento della I_G dovuta all'aumento del campo elettrico laterale, anche a bassa V_G. Questo è dovuto all'accoppiamento tra FG e Drain, che è maggiore in dispositivi più corti: è da notare che la corrente di Gate nella regione di iniezione esibisce una dipendenza decrescente da V_{FG} al ridursi della lunghezza di canale. Questo ha origine dall'aumento del campo elettrico laterale $\varepsilon(x, \bar{y}(x))$ per dispositivi più corti: la riduzione di questo campo, dovuta all'incremento della V_{FG}, si noterà solo ad elevati valori di V_G.

2.1.2 Fowler-Nordheim Tunneling

L'idea dell'effetto tunnel attraverso una barriera di potenziale si applica bene alla struttura dell'ossido sottile nel MOS. La figura 2.6 mostra il diagramma del potenziale in un MOS dove applichiamo un segnale negativo all'elettrodo del substrato di silicio drogato p.

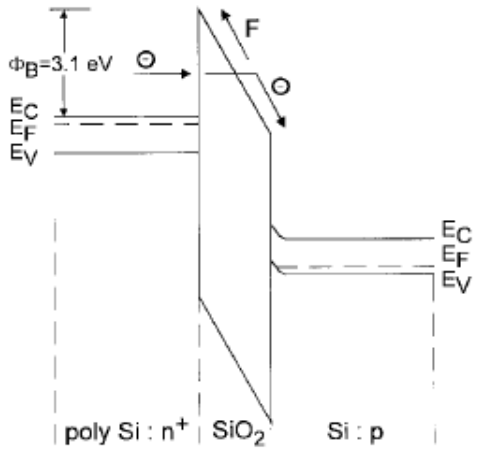

Figura 2.6: I_G in funzione del V_{FG} con L come parametro

La probabilità di effetto tunnel di un elettrone dipende principalmente dalla forma, dall'altezza e dallo spessore della barriera.
Usando il modello dell'elettrone libero e l'approssimazione di Wentzel-Kramers-Brillouin per le probabilità dell'effetto tunnel, otteniamo la seguente espressione per la densità di corrente

$$J = \frac{q^3 F^2}{16\pi^2 h^2 \Phi_B} exp[-4(2m_{ox}^*)^{1/2}\Phi_B^{3/2}/3\hbar qF] \tag{2.19}$$

dove Φ_B è l'altezza della barriera di potenziale, m_{ox}^* è la massa effettiva dell'elettrone nella banda proibita del dielettrico, h è la costante di Planck, q è la carica dell'elettrone, ed F è il campo elettrico attraverso l'ossido.

Nella Figura 2.7 possiamo notare la dipendenza della corrente dal voltaggio: dal momento che il campo è, approssimativamente, il voltaggio applicato diviso lo spessore dell'ossido, una riduzione dello spessore dell'ossido senza una riduzione proporzionale del voltaggio applicato produce un rapido aumento

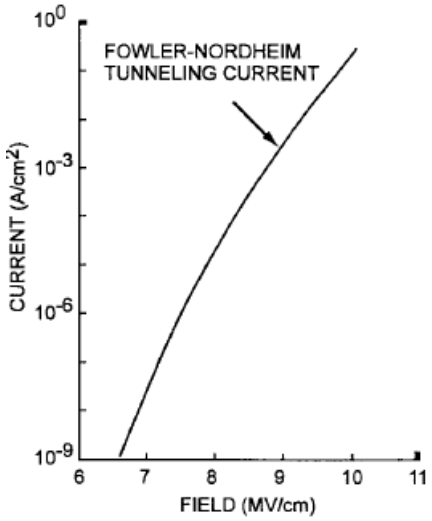

Figura 2.7: $log(J)$ in un FGMOS in funzione di F

della corrente di tunnel. Con un ossido abbastanza spesso (20-30 nm) bisogna applicare un alto voltaggio (20-30 V) per avere un'apprezzabile corrente di tunnel. Con ossidi più sottili, la stessa corrente può essere ottenuta con voltaggi più bassi.
Nei dispositivi attuali è stato scelto uno spessore ottimo (circa 10 nm): essi usano il fenomeno dell'effetto tunnel per avere un compromesso tra i vincoli delle prestazioni (velocità di programmazione, consumo di potenza), per cui ci sarebbe bisogno di ossidi sottili, e le problematiche dell'affidabilità, che vorrebbero ossidi spessi.
Sempre dalla Figura 2.7 è evidente come con un campo di 7 MV/cm la densità di corrente sia circa 10^{-8} A/cm^2, mentre con un campo di 10 MV/cm diventi 10^{-1} A/cm^2, con una variazione di ben 7 ordini di grandezza nel modulo della corrente di tunnel: con un campo di poco superiore potremmo avere anche una differenza di 12 ordini di grandezza.
L'effetto tunnel è usato soprattutto nelle EEPROM, per tre motivi in particolare: è un effetto puramente elettrico; il livello della corrente coinvolta è relativamente basso e permette di generare internamente al chip ogni voltaggio richiesto tramite una pompa di carica; permette al tempo di programmazione ($<$ 1 ms) di essere di 12 ordini di grandezza inferiore a quello di conservazione del dato ($>$ 10 y), fattore fondamentale in qualsiasi memoria non volatile.
Dall'altro lato, la dipendenza esponenziale della corrente di tunnel dal cam-

po elettrico dell'ossido causa alcuni problemi critici nel processo di controllo del dato: ad esempio, una minima variazione dello spessore dell'ossido tra le celle di un array di memoria produce una grande differenza tra le correnti di programmazione e cancellazione, portando ad una distribuzione poco omogenea delle tensioni di soglia in entrambi gli stati logici. Si avrebbe dunque bisogno di un sistema di controllo piuttosto efficiente.
I difetti nell'ossido (la cui densità aumenta con lo spessore dello stesso) devono essere evitati per controllare le caratteristiche di programmazione/cancellazione e per avere una buona affidabilità. In ogni caso, frequenti cicli di operazioni portano ad un conseguente incremento della carica intrappolata nell'ossido: questo influisce sull'altezza della barriera di potenziale, che diventa più bassa nel caso di carica positiva, o più alta nel caso di carica negativa, e quindi aumentando o decrementando le correnti di tunnel.
Sebbene la classica espressione per la corrente di FN (2.19) sia un buon compromesso coi dati sperimentali, molti particolari sono stati finora sottovalutati: la dipendenza del fenomeno dalla temperatura, gli effetti quantici all'interfaccia, l'influenza del "band bending" all'interfaccia del Si/SiO_2, e la caduta di voltaggio nel silicio, il fatto che le corrette statistiche per gli elettroni non siano maxwelliane ma di Fermi-Dirac. Queste caratteristiche sono di fondamentale importanza nella simulazione del dispositivo sia per sviluppare un modello generale per l'iniezione del tunnel che per avere una profonda conoscenza dell'influenza dello scaling nelle prestazioni del dispositivo.
Prima di tutto, la teoria classica parte dal presupposto che gli elettroni, così come le lacune, sulla superficie del conduttore siano trattati come un gas a tre dimensioni formato da particelle libere con una distribuzione di Boltzmann dell'energia. Ma quando la superficie del silicio è in inversione o in accumulo, queste particelle sono confinate in una stretta buca di potenziale, e quindi le leggi della meccanica quantistica richiedono il loro moto perpendicolare all'interfaccia da quantizzare.
Quindi è corretto trattarli come un gas a due dimensioni secondo la meccanica quantistica. I principali risultati di questo sono la scoperta della dipendenza dell'altezza della barriera dal voltaggio e il fatto che il campo dell'ossido sia inferiore al classico grazie alla maggiore perdita di potenziale nel substrato di silicio.
Possiamo riscrivere la (2.19) in forma più semplice come

$$J = A \cdot F^2 exp[-B/F] \tag{2.20}$$

e usare A e B come funzioni del campo elettrico, includendo gli effetti quantici. Questo approccio si è rivelato abbastanza soddisfacente in molti casi ma

può portare a differenti valori di A o B a seconda dell'elettrodo di iniezione e della polarizzazione del dispositivo.

2.2 Operazioni sulla cella

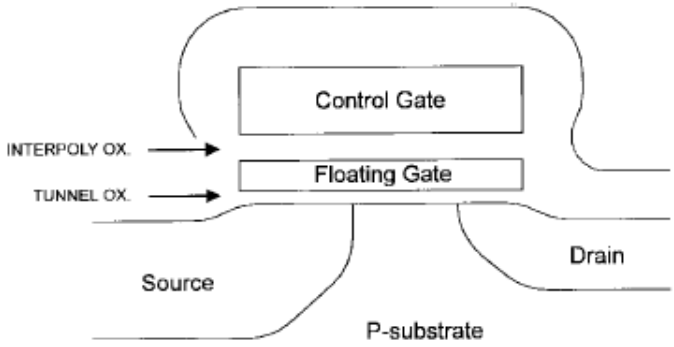

Figura 2.8: Sezione di una cella Flash

Nella Figura 2.8 è mostrata una cella Flash standard, presentata per la prima volta dalla Intel nel 1988 col nome di ETOX (EPROM Tunnel Oxide): vi sono sostanziali differenze da una EPROM, sebbene derivi da essa.
Prima di tutto, l'ossido tra il substrato e il Floating Gate è piuttosto sottile (10 nm), in modo che, all'applicazione di un elevato voltaggio al Source mentre il Gate è a massa, si formi un elevato campo elettrico nell'ossido, permettendo l'effetto tunnel dal Floating Gate al Source. Queste condizioni del segnale sono molto vicine alla rottura della giunzione Source-Bulk: la diffusione di Source è quindi realizzata in maniera diversa da quella di Drain, che non ha a che fare con tali problematiche. Per realizzare questa diversità viene aggiunta una maschera diversa durante la realizzazione del dispositivo: la cella risulta quindi non simmetrica.
La lettura, la programmazione e la cancellazione dei dati sono schematizzate nella Figura 2.9: bisogna in particolare prestare attenzione alla cancellazione, l'operazione più critica.

- *Programmazione*: Viene usata la HEI per caricare il Floating Gate, e dunque cambiare la tensione di soglia del FGMOS. La programmazione è ottenuta applicando simultaneamente un impulso al Control Gate e al Drain quando il Source è a massa: questa operazione può essere effettuata selettivamente applicando l'impulso alla Word Line, a cui sono collegati i Control Gate, e caricando opportunamente la Bit Line, che connette i Drain.

	SOURCE	CONTROL GATE	DRAIN
READ	GND	V_{cc}	V_{read}
PROGRAM	GND	V_{pp}	V_{dd}
ERASE	V_{pp}	GND	FLOAT

Figura 2.9: Segnali tipici durante le operazioni per una cella Flash

Gli elettroni eccitati sono iniettati nel Floating Gate, e la tensione di soglia dei transistor selezionati per l'operazione aumenta: questo aumento varia a seconda della durata dell'impulso di programmazione. Per avere una variazione di voltaggio di circa 3 - 3.5 V, viene usualmente applicato un impulso della durata compresa tra 1 e 10 μs.
Il comportamento si può studiare dalla figura 2.10, in particolare nella curva corrispondente ad $L_{eff} = 0.6$: inizialmente notiamo una rapida variazione della V_T. Successivamente, con il potenziale del Floating Gate che scende al di sotto del potenziale di Drain, V_T satura. Il campo elettrico nell'ossido di tunnel vicino al Drain si inverte, e l'iniezione di elettroni nel Floating Gate diventa molto più difficile. La variazione della tensione di soglia intrinseca non dipende visibilmente dalla lunghezza di canale, ma principalmente dal rapporto di accoppiamento, ovvero la sovrapposizione tra il Floating Gate ed il Control Gate nell'ossido di campo. Possiamo anche notare che la soglia intrinse-

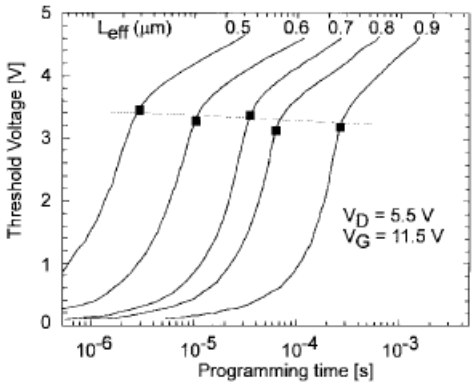

Figura 2.10: Curve di programmazione con diverse lunghezze

ca dipende linearmente dal voltaggio di Drain (Figura 2.11). Anche la temperatura influenza la velocità di programmazione: un'alta temperatura riduce il numero di elettroni eccitati disponibili per l'iniezione nel Floating Gate, ritardando quindi le caratteristiche di programmazione.

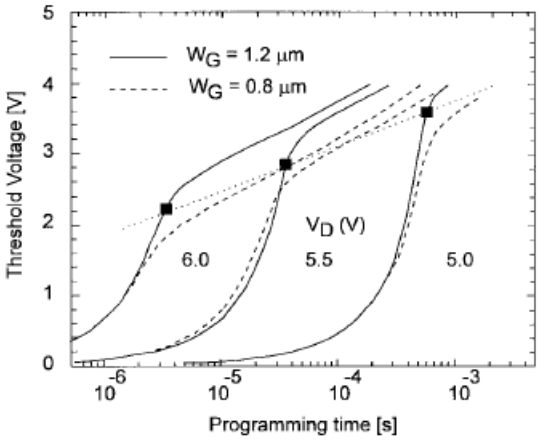

Figura 2.11: Curve di programmazione con diversi accoppiamenti

- *Cancellazione*: Questa operazione richiede un alto voltaggio da applicare al Source (comune a tutti i transistor dell'array) con i Control Gate a massa (tramite la WL) e i Drain flottanti (tramite la BL). Prima di applicare l'impulso di cancellazione, tutte le celle dell'array sono programmate per partire dalla stessa tensione di soglia: di seguito viene applicato l'impulso di durata controllata.
  La variazione della tensione di soglia dipende dal voltaggio del Source: inoltre, per ogni Volt di riduzione del voltaggio di Source, il tempo di cancellazione cresce di un ordine di grandezza, come si nota in Figura 2.12 La tensione di soglia dipende anche dallo spessore dell'ossido (Figura 2.13): dalla Figura uno potrebbe dire che, dopo la cancellazione elettrica, celle con lo stesso spessore dell'ossido ma diversa tensione di soglia iniziale raggiungano la stessa V_T dopo l'operazione di cancellazione. Dal momento che i transistor nell'array hanno diversi spessori dell'ossido di Gate, e il meccanismo di cancellazione non è auto-limitante, dopo un impulso di cancellazione possiamo avere "bit tipici" e "bit a cancellazione rapida". (Figura 2.14) In seguito vengono lette tutte le celle dell'array per controllare se sono state cancellate correttamente: se ciò non è accaduto, vi è un altro ciclo di cancellazione e lettura. Alla

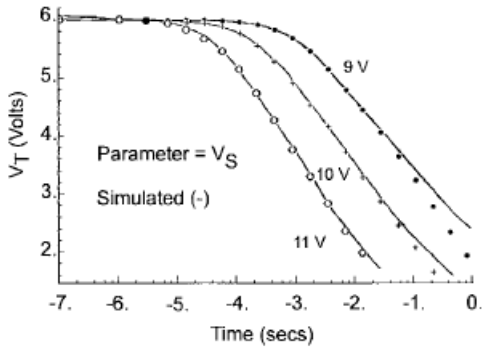

Figura 2.12: Curve di cancellazione con diversi V_S

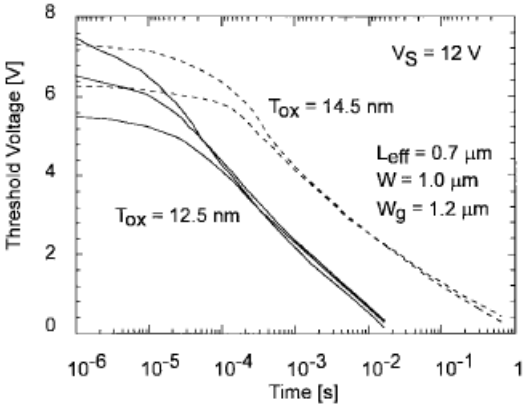

Figura 2.13: Curve di cancellazione con diversi spessori

fine di questa procedura, le celle non avranno la stessa V_T, ma essa avrà una distribuzione gaussiana.
Questa operazione di cancellazione elettrica avviene principalmente attraverso il FN tunneling di cariche dal Floating Gate al Source: per avere una giunzione in grado di sostenere gli alti voltaggi applicati senza rompersi, la giunzione di Source deve essere progettata con precisione, e viene aggiunta al processo una nuova maschera (In Figura 2.15 un dettaglio della giunzione).
Un forte campo elettrico attraverso l'ossido di tunnel porta ad un forte campo anche all'interfaccia del silicio, e questo può significare forti correnti di perdita dovuti al band-to-band tunneling (BBT) o addirittura alla rottura della giunzione Source-Substrato.
Se il "band bending" è superiore alla banda proibita del semiconduttore,

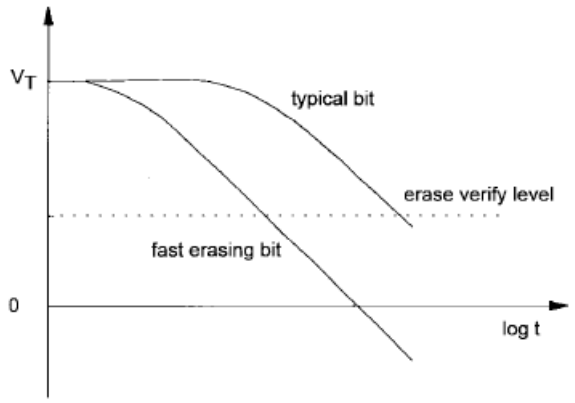

Fig. 19. Erase curves of a "typical bit" and of a "fast erasing bit" in a Flash array [49].

Figura 2.14: Curve di cancellazione per i due tipi di bit

e il campo elettrico sulla superficie è superiore ad 1 MV/cm, l'effetto tunnel degli elettroni dalla banda di valenza a quella di conduzione diventa significativo, e le lacune sono lasciate nella banda di valenza. Gli elettroni sono raccolti dal terminale di Source; le lacune sono iniettate nel Substrato, generandovi così una corrente di perdita, che dipende solo dal campo elettrico verticale nell'ossido, ovvero dalla differenza di potenziale tra Source e Gate. Il campo laterale non permette la generazione di uno strato di inversione nell'interfaccia n^+-Si/SiO_2 e porta ad una regione di spazio di svuotamento, senza alcun portatore libero. Quando il potenziale al Source è sufficientemente alto, l'impatto tra i portatori diventa significativo e contribuisce alle correnti di perdita, dando inizio alla rottura. Il voltaggio minimo per dare inizio al BBT diminuisce alla riduzione dello spessore dell'ossido, e questo è uno dei principali problemi nello scaling. Le lacune generate possono guadagnare energia per essere iniettate nell'ossido dove sono intrappolate nell'interfaccia del Si/SiO_2.
La rottura del Source è uno dei maggiori limiti della riduzione del tempo di cancellazione, dal momento che con un aumento della tensione al Source esso diminuisce: una soluzione è ottenuta ottimizzando il profilo della giunzione di Source ad una più graduale per ridurne il campo elettrico e conseguentemente la corrente di substrato di qualche ordine di grandezza.
In una cella Flash convenzionale a due tensioni, dove oltre alla consueta V_{CC} (3-5 V) abbiamo disponibile un'altra tensione, elevata (circa 12

V), V_{pp}, la cancellazione è ottenuta applicando un'alta tensione positiva alla regione di Source, mentre la WL (e quindi il Control Gate) è a massa. In una cella a singolo voltaggio invece la mancanza di V_{pp} implica una generazione interna al chip di un voltaggio negativo tramite una pompa di carica: in questo caso la differenza di potenziale necessaria tra il Source e il Control Gate è ottenuta applicando V_{CC} al Source e una tensione negativa V_{GN} al Control Gate. Non importa come questa differenza sia ottenuta, in entrambi i casi il forte campo elettrico nell'ossido tra Floating Gate e Source dà origine ad una corrente di Gate dovuta al FN tunneling: simultaneamente, il campo elettrico nel silicio è responsabile della corrente tra Source e Substrato dovuta al BBT tunneling, funzione della tensione di Source.

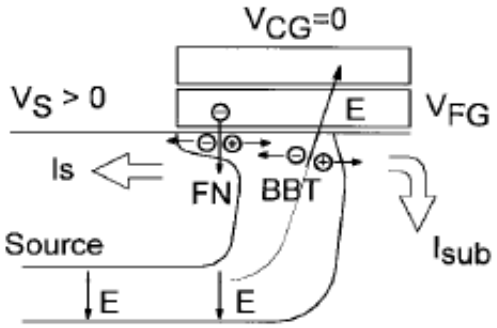

Figura 2.15: Dettaglio della giunzione di Source

- *Lettura*: Il dato salvato in una cella Flash può essere misurato determinando la tensione di soglia del FGMOS. Il modo migliore e più veloce per farlo è leggendo la corrente che scorre nella cella con un segnale fissato al Gate: come si può notare dalla Figura 2.16, i valori logici "1" o "0" presentano la stessa curva di transconduttanza, ma spostata di una determinata quantità - la variazione della tensione di soglia ΔV_T - proporzionale alla carica immagazzinata Q.
  Perciò, quando è immagazzinata una determinata carica C e abbiamo la conseguente variazione di tensione ΔV_T, è possibile fissare un voltaggio di lettura in modo che, se il dato presente è "1", la corrente sia molto alta (sulle decine di μA), mentre nel caso di "0" la corrente sia pressochè nulla. In questo modo è possibile definire uno stato logico "1" da un punto di vista microscopico (nessuna carica immagazzinata) e macroscopico (alta corrente): viceversa, per lo stato logico "0" abbiamo carica immagazzinata e corrente nulla.

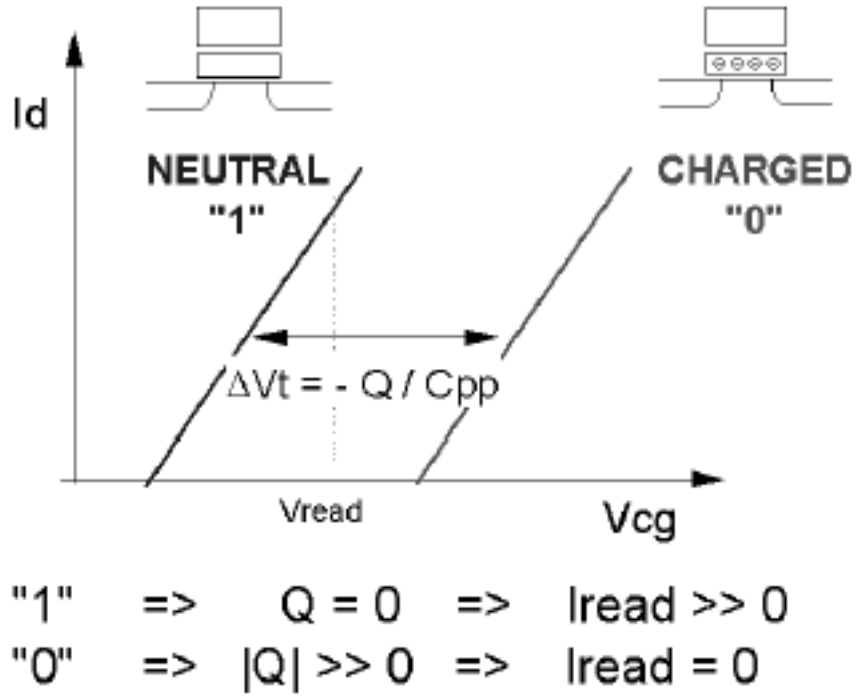

Figura 2.16: Operazione di lettura in un FGMOS

2.3 L'architettura a NAND

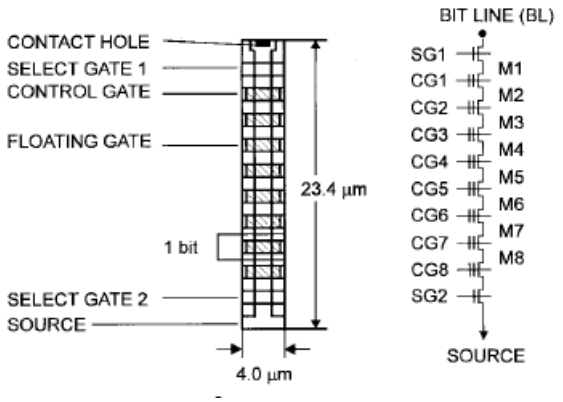

Figura 2.17: Architettura NAND

L'approccio per l'organizzazione di un array di celle in una struttura a NAND consente un grande miglioramento nelle prestazioni e nei risultati rispetto ad altre architetture: l'unita base è costituita da una serie di FGMOS (solitamente 8 o 16) che costituiscono una catena connessa alla Bit Line e alla massa tramite due transistor di selezione (Figura 2.17). Questa organizzazione permette l'eliminazione di tutti i contatti tra le Word Line, che possono essere separate dalle minime regole di layout, riducendo l'area occupata di circa il 40% (Figura 2.18). Per di più, un tipo di memoria che ha come un'unità base una cella che immagazzina un solo bit è più vicina alla memoria ideale con l'accesso in parallelo: permette di programmare anche pagine da 256 byte, avendo un'ottima versatilità.

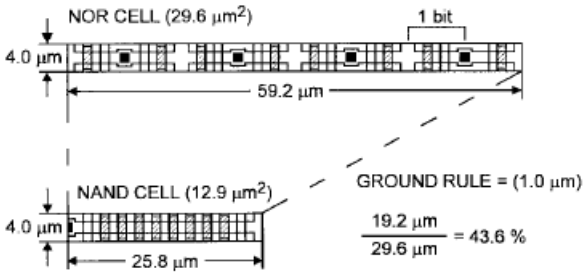

Figura 2.18: Confronto tra le architetture

Nella Figura 2.19 viene mostrata una sezione di un blocco elementare da 8

bit di una memoria Flash da 4 MB organizzata a NAND insieme ai circuiti periferici.

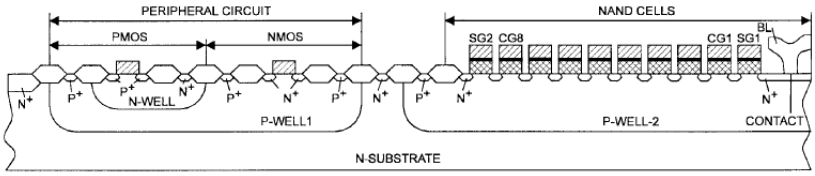

Figura 2.19: Cella NAND e circuiti periferici

I voltaggi di cancellazione sono circa 20 V per il n-substrato, i p-well2, il Drain ed il Source, mentre è 0 V per il Control Gate del transistor selezionato. Questi segnali inducono l'effetto tunnel attraverso la p-well2, portando ad una bassa soglia negativa. Non c'è differenza di potenziale tra Drain e p-well, quindi non c'è breakdown nella giunzione. I voltaggi di programmazione sono 18 V sul Control Gate selezionato, 10 V per gli altri Gate, il p-well2 invece è a massa.
I transistor di selezione ricevono il segnale per connettere la catena alla Bit Line e isolarla dalla massa: se deve essere salvato uno "0", la BL è a massa. Source e Drain sono a massa, e solo il transistor selezionato ha un campo elettrico nell'ossido tale da indurre l'iniezione di elettroni dal substrato al Floating Gate, aumentando la tensione di soglia al livello alto. Se invece deve essere salvato un "1", sulla BL viene portato un segnale a 10 V: non c'è effetto tunnel e la soglia è negativa.
L'operazione di lettura è effettuata applicando 5 V a tutti i Control Gate eccetto quello selezionato, che è a massa: i transistor di selezione ora conducono, e connettono la catena alla BL e alla massa: la BL è precaricata e di conseguenza, se il dato da leggere è "1" e il transistor selezionato conduce, essa viene scaricata a massa; se invece il dato da leggere è "0", il transistor è spento e la BL mantiene la carica.
Se l'array ha un'architettura a NAND, il controllo delle tensioni di soglia è un fattore critico, poiché sono ammesse solo piccole variazioni rispetto al valore nominale. Negli array a NAND sia il meccanismo di programmazione che quello di cancellazione avvengono tramite l'effetto tunnel degli elettroni: dal momento che l'effetto tunnel è più efficiente della HEI, le correnti sono inferiori e le differenti tensioni di alimentazione possono essere generate internamente con pompe di carica implementate nello stesso die. Questa architettura è preferita per le applicazioni Flash per la grande densità delle celle.

Il tempo di accesso alle pagine è circa 80 ns, e anche la programmazione può essere effettuata sulla pagina, riducendo il tempo di programmazione a circa 300 μs per pagina: la cancellazione è invece veloce e richiede 6 ms per blocco e circa 10 ms per chip.

Capitolo 3

Problematiche

3.1 Affidabilità

La cella Flash eredita molte problematiche, data la struttura e il metodo di funzionamento simili, dalle celle EPROM ed EEPROM: gran parte di esse sono dovute al continuo aumento della densità delle memorie, che influisce principalmente sulla velocità e sull'affidabilità (con affidabilità si intende la qualità di mantenere più a lungo possibile nel tempo le caratteristiche fondamentali e garantite del prodotto appena uscito dalla fabbrica).
Dalla progettazione alla realizzazione di una memoria, si può incorrere in un gran numero di problemi, integrando milioni di celle su un unico array: la nostra conoscenza sull'argomento, e quindi la fiducia nelle potenzialità e nella sicurezza di una memoria Flash, cresce a mano a mano che aumenta la comprensione dei fenomeni di errore interni alla cella.
Le possibilità di test permettono il riscontro di difetti latenti sin dal wafer, i quali portano a errori a livello di cella come ad esempio la perdita del dato, i difetti dell'ossido, disturbi di programmazione e altre problematiche comuni alle memorie Flash, che ciononostante rimangono comunque una delle memorie non volatili più affidabili sul mercato.
Tra memorie a NAND e memorie a NOR, comunque, possiamo riscontrare delle differenze, principalmente per il fatto che le NAND contengono un maggior numero di blocchi mal funzionanti: nonostante il costruttore indichi un numero minimo di celle correttamente funzionanti, la locazione di quel-

le difettose deve essere scoperta dall'acquirente tramite il successivo utilizzo della memoria stessa. Non è tutto, poiché anche con l'eccessivo utilizzo alcune celle potrebbero diventare non funzionanti, ed anche il loro mancato funzionamento si riscontra solamente durante l'esecuzione.
Uno dei principali punti di forza di questa architettura, comunque, rimane il limitato consumo di potenza: essendo dispositivi a Metal-Oxide-Semiconductor, consumano poca energia nei periodi in cui non sono utilizzati, e anche durante le operazioni stesse. E, come ogni dispositivo a stato solido, le memorie Flash hanno un'ottima resistenza ad urti e vibrazioni, potendo lavorare anche in un ampio raggio di temperature.

3.2 Principali fattori di errore

3.2.1 Distribuzione della tensione di soglia

Avendo a che fare con un gran numero di celle, anche nell'ordine del milione, bisogna comprendere a pieno come il fattore della dispersione della tensione di soglia influisca sul corretto funzionamento del dispositivo, poiché essa varia a seconda dell'operazione appena effettuata e deve essere controllata per non essere eccessivamente distribuita su valori troppo lontani da quello prestabilito: il modo migliore per comprenderlo è dunque studiare il suo comportamento su tutto l'array in seguito a diverse operazioni specifiche, in particolare la cancellazione UV, la programmazione HEI e la cancellazione per FN tunneling.

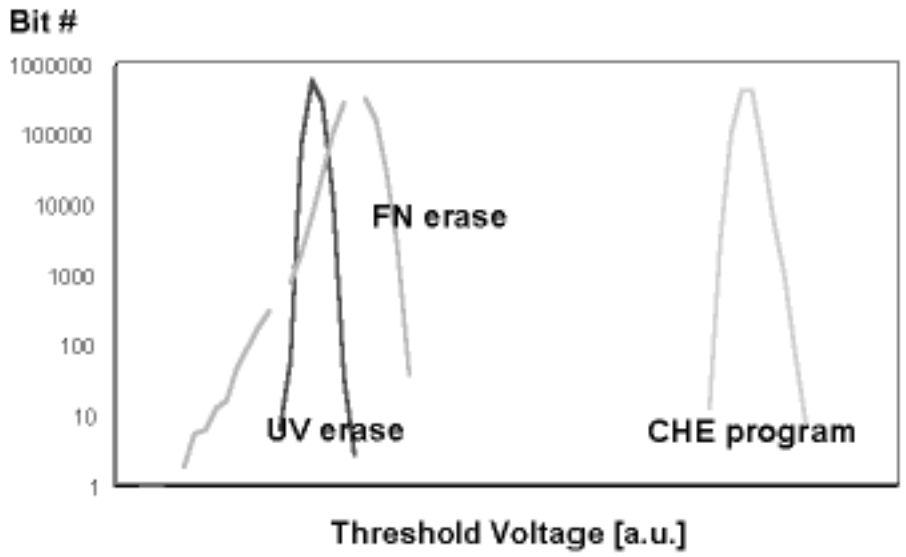

Figura 3.1: Distribuzione della tensione di soglia

Come possiamo notare dalla Figura 3.1, dopo la cancellazione UV abbiamo una distribuzione piuttosto stretta e simmetrica: uno studio approfondito ci rivela che essa è una distribuzione gaussiana, dipendente in modo particolare da variazioni casuali dello spessore dell'ossido, dalle dimensioni critiche e dal drogaggio.
Dopo la programmazione CHE, la distribuzione è ancora simmetrica, sebbene molto più ampia: ciò avviene principalmente perchè i fattori che han

contribuito alla dispersione della cancellazione UV si riflettono anche su questa programmazione, dando vita a quest'amipa distribuzione.
Dopo la cancellazione elettrica, invece, otteniamo una distribuzione ancora più sparsa e soprattutto non simmetrica: il culmine di essa continua ad essere una distribuzione gaussiana con una deviazione sicuramente maggiore rispetto a quella di programmazione (e le celle che soddisfano questo requisito vengono definite celle "normali"), ma ci ritroviamo con una coda di celle, per basse V_T, che differisce in maniera sostanziale dalla tensione di soglia media voluta, seguendo piuttosto una legge esponenziale, e che quindi si cancellano molto più velocemente rispetto alle normali celle (e vengono chiamate celle "coda").
La dispersione delle celle normali è stata ampiamente studiata, ed è stata trovata la dipendenza della loro distribuzione dalle variazioni nel fattore di accoppiamento: per le celle coda, invece, elemento chiave in una memoria, lo studio del funzionamento risulta più complicato. Si ritiene però che il loro particolare comportamento sia dovuto a difetti nell'elettrodo di iniezione o a cariche presenti nell'ossido. Per il fatto che esse si cancellano molto più velocemente rispetto alle normali celle, sarebbe comune pensarle come difettose: ma il loro numero è talmente elevato che non sono semplicemente classificabili come difetti di fabbricazione.
Lo studio delle celle coda ha portato a vari modelli per spiegare questo fenomeno: ad esempio, una distribuzione nella struttura policristallina del Floating Gate, con una variazione nell'altezza della barriera tra due cristalli, darebbe luogo ad un aumento locale della barriera del tunnel. Un altro modello spiega la presenza di queste celle coda con la distribuzione casuale di cariche positive nell'ossido di tunnel: esso è supportato dalla conoscenza di trappole donatrici presenti nell'ossido che mostrano l'incremento della densità di corrente nel tunnel causata dalle cariche positive nelle vicinanze dell'elettrodo di iniezione.
Indipendentemente dal modello, si può dimostrare che la coda di celle con dipendenza esponenziale è dovuta a imperfezioni strutturali (difetti intrinseci) e può essere minimizzata con processi di ottimizzazione, ma la sua eliminazione non è possibile: le memorie Flash devono essere progettate tenendo conto di queste imperfezioni.

3.2.2 Disturbi di programmazione

Con "disturbo di programmazione" si intende la corruzione del dato salvato in una cella dovuta allo stress a cui è sottoposta anche nel momento in cui

si stanno programmando altre celle dell'array di memoria.

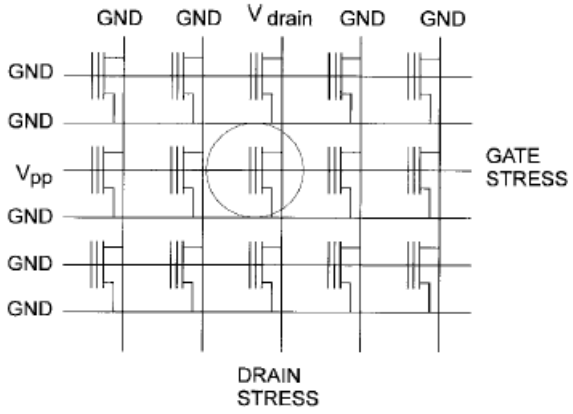

Figura 3.2: Array Flash

Nella Figura 3.2 abbiamo lo schema di un'array di memoria Flash, in cui è evidenziata la cella che vogliamo programmare, e le corrispondenti righe e colonne che vengono influenzate durante la scrittura della stessa per effetto dei disturbi: i due differenti tipi di disturbo vengono chiamati di Gate per le righe, di Drain per le colonne.
Nel caso del disturbo di riga, se un'elevata tensione è applicata alla WL, tutte le altre celle devono sopportare lo stress senza la perdita del dato: essa può accadere per perdite nell'ossido di gate o nel dielettrico interpoly. Supponiamo quindi di applicare sulla WL selezionata una tensione di circa 12 V e, per la generazione di elettroni eccitati per effettuare la programmazione, venga applicato alla BL un segnale di 5-7 V: i Gate degli altri transistor potrebbero quindi essere sotto stress, e potremmo avere effetto tunnel degli elettroni dal Floating Gate al Control Gate attraverso l'ossido interpoly in tutte le celle programmate, ovvero quelle in cui abbiamo cariche nel FG, poiché è applicata una tensione di 12 V al Gate e di 0 V sia al Source che al Drain. Questo disturbo, che porta alla perdita di carica e riduce quindi la tensione di soglia, è chiamato "dc-erasing"; ma oltre all'effetto in queste celle, l'effetto tunnel potrebbe anche esserci nelle celle non programmate, dal Substrato al Floating Gate, ovvero nei transistor "vuoti", portando all'acquisizione di carica e alla modifica della tensione di soglia, effetto chiamato "dc-programming". Entrambi questi effetti sono presenti anche durante le operazioni di lettura.
Quando invece è applicato un alto voltaggio (5-7 V) alla BL, esso può stressare i Drain di tutti i transistor nella stessa colonna di quello selezionato per

l'operazione in corso: avendo in comune la stessa BL, gli elettroni subiscono l'effetto tunnel dal FG al Drain, attraverso l'ossido di Gate: possono anche generarsi lacune dovute agli impatti nel substrato, e conseguentemente esse vengono iniettate nel Floating Gate. Questo disturbo è chiamato "drain disturb" e causa perdite di carica e un decremento della tensione di soglia: esso può avvenire anche in caso di cicli di scrittura troppo lunghi, ed è spesso usato come indicatore della qualità dell'ossido di Gate.
Quindi questi disturbi diventano rilevanti soprattutto dopo ripetute operazioni di lettura e scrittura, ad esempio quando esse devono essere effettuate per un'intera riga o per un'intera colonna.

3.2.3 Conservazione dei dati

Ad una memoria non volatile è richiesta la capacità di mantenere i dati anche per una decina d'anni, resistendo a cicli di scrittura e cancellazione ed essendo sottoposta a forti tensioni e correnti attraverso gli ossidi sottili: tutto ciò potrebbe portare alla perdita di carica immagazzinata nel Floating Gate, basti pensare che ad una tensione di soglia abbastanza comune, 2 V, siamo in presenza di un numero di elettroni nell'ordine di 10^3 - 10^4, ed una perdita di un semplice 20% di questa quantità potrebbe portare ad una lettura erronea della cella, e dunque ad una perdita del dato.
Possiamo dividere i difetti che portano a queste perdite in due categorie: intrinseci, dovuti ai difetti interni del dispositivi, ed estrinseci, causati dai meccanismi fisici che sono utilizzati per la programmazione e per la cancellazione.

- Tra i difetti intrinseci che possiamo trovare abbiamo ad esempio l'emissione di elettroni assistita da campo, e consiste nel moto di elettroni immagazzinati nel FG di una cella programmata, che migrano verso l'interfaccia con l'ossido e da qui subiscono l'effetto tunnel per arrivare al substrato, causando la perdita di carica. Se invece la cella ha la tensione di soglia di livello basso, si verifica l'iniezione opposta. Questa corrente di perdita dipende dal coefficiente di accoppiamento tra FG e CG α_G e dal livello di stress: la probabilità che un'elettrone passi attraverso l'ossido dipende dalla differenza di potenziale tra FG e substrato, e il potenziale di FG dipende dal CG tramite α_G. La carica Q nel FG scende con il diminuire di α_G o con l'aumentare del livello di stress: la corrente di perdita dipende esponenzialmente dal campo

elettrico secondo

$$E = \frac{Q}{2\epsilon_{ox}\sigma} \tag{3.1}$$

con ϵ_{ox} la costante dielettrica del silicio e σ l'area del FG.
Un altro difetto che porta alla perdita di carica, l'emissione termoionica, è un meccanismo di emissione di portatori al di sopra della barriera di potenziale. A temperatura ambiente, il fenomeno è trascurabile, ma ad alte temperature inizia a diventare rilevante.

- Tra le cause estrinseche ci sono ad esempio i difetti dell'ossido, che possono causare perdita o aumento della carica: se la cella è programmata, il FG ha un potenziale negativo data la carica immagazzinata, e questo potenziale induce un campo elettrico nell'ossido che circonda lo stesso FG. Nell'ossido sottile, questi campi elettrici possono diventare anche piuttosto forti, e i difetti nell'ossido possono indurre cammini conduttivi che arrivando a programmare la cella; se la cella invece è cancellata e possiede una carica positiva, allora il campo elettrico induce un guadagno di carica.
  La contaminazione ionica è un grande problema per ogni memoria non volatile: gli ioni, solitamente positivi, vengono attratti dal FG caricato negativamente, inducendo ad una perdita di carica.

3.2.4 Resistenza a programmazione e scrittura

La garanzia di durata di una memoria Flash è di circa 10^5 cicli di programmazione e cancellazione: essi però possono causare una degradazione, piuttosto uniforme, delle prestazioni della cella, dovuta principalmente alla degradazione dell'ossido di tunnel.

Come notiamo dalla Figura 3.3, la variazione della tensione di soglia in programmazione e in cancellazione ci danno una misura dell'invecchiamento dell'ossido di tunnel: nelle vere memorie Flash vengono usati algoritmi per prevenire questa degradazione, che però portano ad un incremento del tempo di programmazione e cancellazione. In particolare, la riduzione della tensione di soglia in programmazione è dovuta alla generazione di trappole nell'ossido, meccanismo che porta alla degradazione degli elettroni eccitati.
Il cambiamento della tensione di soglia in cancellazione, invece, riflette le dinamiche di una carica prefissata nell'ossido di tunnel in funzione della carica iniettata: la degradazione iniziale della V_T è dovuta ad un accumulo di cariche positive che aumentano l'efficienza dell'effetto tunnel, mentre l'incremento a

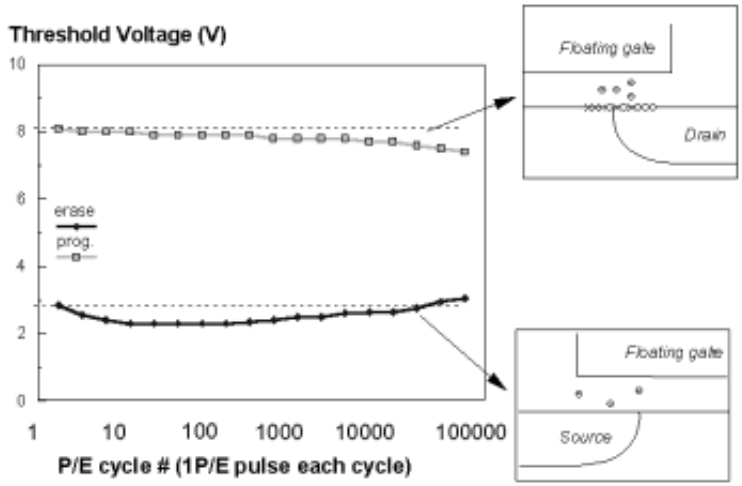

Figura 3.3: Tensione di soglia in funzione del numero di cicli

lungo termine della V_T è dovuto alla generazione di trappole negative. Tutte queste problematiche possono essere ridotte ottimizzando il processo utilizzato per l'ossido di tunnel: una volta che il processo e la produzione garantiscono una determinata resistenza, non dovrebbero esserci grandi variazioni tra i lotti prodotti. I problemi di resistenza, infatti, vengono principalmente dalle singole celle, presentando esse stesse difetti di conservazione dopo cicli di programmazione/cancellazione. Un elevato stress sull'ossido sottile aumenta la densità di corrente a basso campo elettrico: questo eccesso di corrente, che causa una deviazione significativa dalle curve FN caratteristiche, è conosciuto come "stress-induced leakage current" (SILC), attribuito ai difetti dell'ossido e alle trappole che aiutano l'effetto tunnel. I principali fattori da cui il SILC dipende sono il campo dovuto a questo stress, la quantità di carica iniettata durante lo "stress", e lo spessore dell'ossido. Per condizioni di stress fissate, la corrente di perdita aumenta significativamente con spessori di ossido minori. Lo studio della resistenza va quindi effettuato su interi array, e non su singole celle, in particolare sulle celle coda, e questo problema è il principale limite allo scaling dello spessore dell'ossido di tunnel: sotto 8-9 nm, il numero di celle difettose diventa troppo alto, e nemmeno tecniche di correzione dell'errore sofisticate possono correggere questo problema.

Capitolo 4

Applicazioni e utilizzo

Nel campo dell'elettronica, il mercato delle memorie può essere principalmente diviso in due categorie: le memorie portatili a bassa capacità e le memorie ad alte capacità. Le Flash, ovviamente, dominano il mercato delle basse capacità e continueranno a farlo in futuro, principalmente per i costi contenuti; d'altro canto però gli HDD sono il principale prodotto venduto nel campo delle alte capacità, dal momento che il requisito principale per questo tipo di memorie è il costo del prodotto per GB, requisito in cui gli HDD sono tuttora superiori rispetto alle memorie Flash.
Per il mercato di notebook e desktop, infatti, la richiesta di alte capacità continua a rimanere il requisito primario, dal momento che l'immagazzinamento di risorse quali foto, musica e video continua ad espandersi e richiederà sistemi di memoria sempre maggiori: ciononostante, c'è un leggero ma costante aumento del segmento di mercato dove non sono richieste alte capacità, ad esempio nel caso dove i clienti vogliano mantenere solo una "cache" dei dati frequentemente usati e dei programmi. In questo caso delle SSD o delle Flash integrate potrebbero bastare.
Le memorie SSD sono un'altra forma di memoria non volatile che a poco a poco sta rimpiazzando anche gli HDD: esse usano al 100% l'architettura Flash a NAND per immagazzinare il dato, e poichè appaiono al sistema proprio come un HDD non vi è alcun bisogno di installazione di nuovi driver o di altri sitemi di interfacciamento.
Le SSD non soffrono dei problemi meccanici e di rotazione tipici degli HDD, e l'accesso a settori casuali è limitata solo dalla latenza di lettura di una pagina di una NAND: dunque le operazioni ad accesso casuale per una SSD possono

essere tra le 10 e le 50 volte più veloci di un HDD; per gli accessi sequenziali invece gli HDD continuano ad essere più rapidi. Purtroppo alcuni problemi intrinseci delle NAND potrebbero comunque rendere le SSD più lente degli HDD in alcuni accessi casuali o sequenziali, come ad esempio la latenza di lettura più lunga o i tempi di cancellazione di blocchi più lenti.
La combinazione di questi differenti fattori porta al fatto che la maggior parte delle SSD indirizzate ai notebook non siano sensibilmente più veloci degli HDD, ma le SSD possono superare molti di questi problemi prestazionali aumentando il parallelismo e aggiungendo una cache di scrittura. Alcune SSD hanno già integrato queste caratteristiche, mostrando significativi miglioramenti seppur con un aumento non indifferente del costo.
L'affidabilità a lungo termine sulle SSD, invece, rimane un problema aperto: pur essendo un apparecchio a stato solido e quindi potenzialmente più robusto di un HDD, il numero limitato di cicli di cancellazione per una Flash potrebbe essere una limitazione.

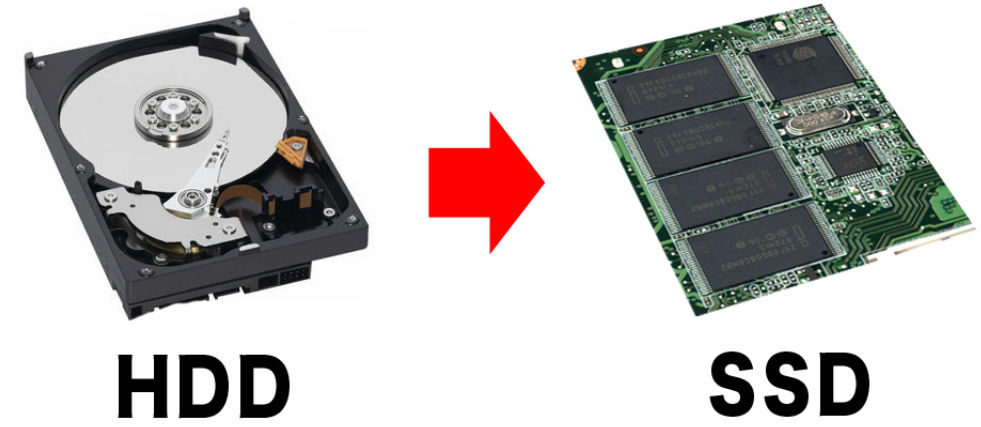

Figura 4.1: HDD ed SSD a confronto

Ma un altro problema relativo agli utilizzi delle memorie Flash nei PC riguarda le possibili scelte che derivano dall'utilizzo della tecnologia: le varie alternative, tra cui HDD ibridi, SSD e memorie turbo, presentano ognuna diversi punti di forza, lasciando all'utente la scelta a seconda del diverso utilizzo che deve fare della sua memoria non volatile. Vi sono quindi campi in cui potrebbe essere richieste le alte prestazioni delle SSD, seppure con esse venga dietro anche un alto prezzo. Ma per gli utilizzi principali di immagazzinamento dati, ovvero dove c'è bisogno di terabyte di capacità, un uso così significativo delle memorie Flash non è economicamente consigliato neppure

nel futuro più prossimo.

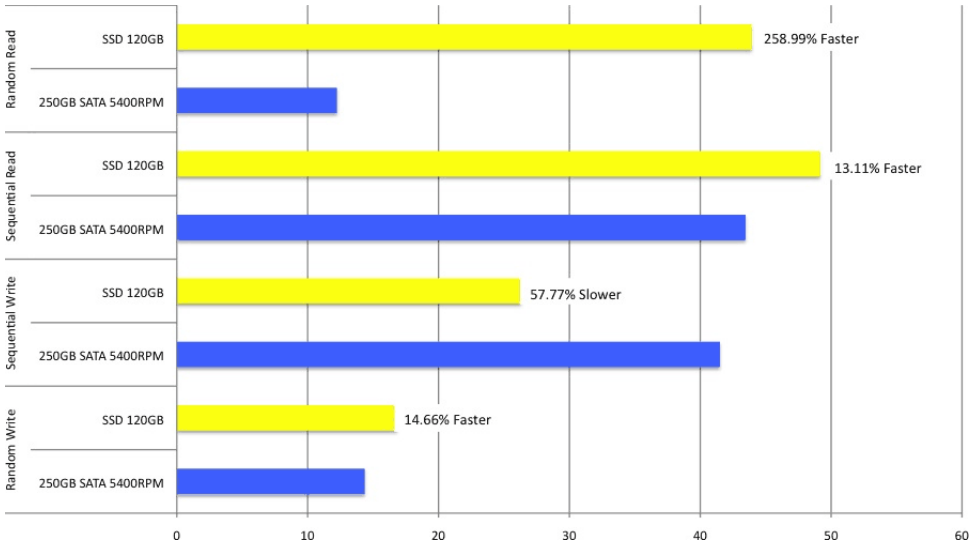

Figura 4.2: Prestazioni a confronto

Un modo, dunque, per combinare le elevate capacità di un HDD e la velocità di una memoria Flash può essere l'utilizzo di quest'ultima come memoria "cache" in aggiunta ad un HDD per prevenire i problemi dovuti alla latenza meccanica del dispositivo, oppure l'utilizzo come buffer per le richieste di accesso ai dati delle CPU.
L'utilizzo più diffuso per la memoria Flash, però, rimane quello delle schede di memoria: pensiamo ad esempio alle SD che inseriamo nel nostro cellulare o nella macchina fotografica, o anche alle Memory Card utilizzate da console casalinghe quali PlayStation. Esistono vari modelli di schede, ognuno con la propria interfaccia e dimensioni specifiche che variano a seconda dell'utilizzo per cui sono predisposte: ad esempio per i telefoni cellulari si necessita di schede dalle dimensioni contenute (μSD o MMC ridotte), mentre per le fotocamere digitali si possono avere schede più grandi che richiedono però una maggiore densità di memoria (CF, SD, MMC). Una problematica che si riscontra in questa applicazione è la capacità delle schede di sopportare l'inserzione e la rimozione, ovvero resistere all'improvvisa mancanza di una fonte di alimentazione senza perdere il dato salvato.

Una scheda è un sistema piccolo e completo dove ogni componente è saldato su una printed circuit board (PCB) e messo in package indipendenti: in questo modo è possibile aggiungere componenti aggiuntivi esterni, come ad

Figura 4.3: Gamma di schede di memoria

esempio dei generatori di segnali di clock o dei generatori di potenza. Ciò però non è attuabile in schede di dimensioni contenute come le μSD, le quali hanno le dimensioni di un die di una memoria a NAND, e vi è una stretta limitazione alla massima densità della scheda, essendo montate direttamente su sottili die. Un'altra problematica è quindi lo stress da assemblaggio, che richiede l'attuazione di altri test sull'efficienza della scheda e quindi a costi maggiori.
I modelli più diffusi di schede di memoria sono quelli da 2 o 4 GB, grazie soprattutto al costo contenuto, ma possiamo avere anche modelli fino a 64 GB (CompactFlash).

Capitolo 5

Conclusioni e sviluppi futuri

Le memorie Flash subiscono un costante confronto con le controparti meccaniche, gli HDD, sotto vari profili: il costo, le prestazioni, l'utilizzo. Abbiamo già ampiamente discusso riguardo le possibili applicazioni della tecnologia Flash, ed esaminiamo dunque gli altri due campi per giungere a delle possibili conclusioni a riguardo.
Sul prezzo influiscono ovviamente i costi alla fabbrica, come ad esempio la manodopera, i componenti, ecc.: bisogna però tener conto del fatto che il prezzo a cui le memorie vengono venute è influenzato da vari fattori, quali il particolare periodo storico, la domanda, la disponibilità, i livelli di integrazione e una possibile sovrapproduzione del prodotto. Per questa analisi prenderemo in esame il caso di integrazione minima, escludendo quindi qualsiasi possibile fattore esterno.
Per gli HDD possiamo notare una grande varietà nei possibili prezzi, dovuta principalmente al fatto che, per la sua fabbricazione, vi è bisogno un numero non indifferente di componenti meccaniche: ciò influisce soprattutto per le memorie a capacità inferiore, poiché ad un fissato costo per GB di memoria le componenti meccaniche rimangono invariate, pesando sostanzialmente sul prezzo di una memoria a bassa capacità. Non è tutto, poiché tra gli HDD abbiamo un ampio range di prezzi, a seconda della forma, della capacità, dell'affidabilità e delle prestazioni, e possiamo dunque avere HDD i cui costi differiscono anche di un ordine di grandezza (x10).
Per una memoria Flash invece il costo medio è dato dal costo per la produzione di un wafer diviso il numero di celle realmente funzionanti: gran parte del costo deriva dai processi e dai macchinari utilizzati: in ogni caso il costo

medio per chip diminuisce a poco a poco ogni anno, dovuto all'aumento delle dimensioni del wafer e ai tempi di produzione sempre più ridotti. Un altro motivo che ha portato alla costante diminuzione del prezzo è lo sviluppo della tecnologia MLC, ovvero a celle con diversi livelli di tensione di soglia in grado di poter assumere quindi più di due valori logici, e che ha portato ad una maggiore densità e ad un maggior numero di bit per cella. Dunque negli ultimi cinque anni abbiamo assistito ad un dimezzamento del costo delle memorie Flash: ciò è però dovuto a eventi unici che molto probabilmente non si ripeteranno, e dunque nel futuro non potremo aspettarci una tale diminuzione del prezzo.
Per i prossimi anni gli analisti hanno previsto una costante diminuzione del prezzo di entrambe le tecnologie, con gli HDD che rimarranno dunque i più convenienti dispositivi confrontando i $/GB.
Riguardo alle prestazioni, il confronto è sicuramente più semplice: le Flash sono più appetibili soprattutto per le operazioni ad accesso e scrittura casuale, mentre per le operazioni sequenziali i problemi delle Flash possono essere risolti con l'utilizzo del parallelismo. Il costante aumento delle prestazioni di dispositivi nuovi quali le SSD sta portando ad una sempre maggior diffusione e ad un continuo sviluppo in questo campo: al giorno d'oggi la presenza di SSD in notebook o Personal Computer non è più un'eccezione, essendo diventata una risorsa affidabile ed estremamente competitiva grazie all'elevata velocità di accesso, a scapito però di una minore affidabilità sul lungo periodo.
Riguardo invece alle schede di memoria, il loro larghissimo utilizzo non è certo poco noto: qualunque dispositivo portatile, come ad esempio un telefono cellulare o una macchina digitale, ne fa costante utilizzo, rendendo le memorie Flash gli assoluti dominatori di questa fetta di mercato, avendo ampiamente soppiantato gli HDD.
Con più di vent'anni di costante utilizzo e miglioramento, dunque, le memorie Flash sono sicuramente un dispositivo largamente diffuso e in grado di soddisfare un'ampia fetta di mercato grazie alla sua grande flessibilità e ai costi contenuti rispetto agli altri generi di memorie non volatili: il loro ingresso ha cambiato radicalmente il mondo delle memorie a semiconduttore, ponendo uno standard con cui ogni futura architettura di immagazzinamento dati dovrà confrontarsi. I continui miglioramenti sia nelle prestazioni che nei costi non fa che aumentare la loro diffusione, rendendo però precluso il mondo delle memorie con bassi rapporti $/GB o in cui vi è bisogno di accessi rapidi e un'affidabilità comprovata nella lunga distanza.
Non sarà inoltre possibile assistere ad un altro sostanziale calo dei costi di questo prodotto, essendo, come precedentemente detto, dovuto ad eventi unici difficilmente ripetibili: le future tecnologie a 3x o 4x MLC dovranno

dimostrare la loro affidabilità e la possibilità di essere scalate anche su litografie di dimensioni inferiori. Vi sono inoltre nuove memorie non volatili che a lungo andare potrebbero soppiantare le memorie Flash proprio nei campi in cui esse risultano fallaci, ovvero la resistenza e la velocità di scrittura: questo è il caso delle già discussione SSD. Recenti sondaggi della Kroll Ontrack, azienda leader del recupero dati da Hard Disk, hanno contribuito a dare un'immagine più chiara della diffusione di questa tecnologia, sempre più utilizzata sia a livello domestico che aziendale. Le SSD infatti risultano sempre più diffuse soprattutto tra i manager, utilizzatori di apparecchi quali tablet o dispositivi mobili, grazie in particolare alla loro maggior velocità di accesso dei dati; secondariamente viene preferito il loro utilizzo per l'affidabilità e per il risparmio energetico.
Va tenuto conto però del fatto che, sempre secondo il sondaggio, la maggior parte degli utilizzatori non è a conoscenza della possibile durata media di una memoria SSD, e i due terzi degli intervistati ritengono le SSD ancora meno sicure rispetto agli HDD, principalmente per una possibile futura difficoltà nel recupero dati dovuta alla presenza di crittografia proprietaria o per l'assenza di garanzie sulla cancellazione sicura dei dati.
Il futuro mercato delle memorie Flash, dunque, appare ancora in ottima salute: nonostante un calo subito nelle vendite nell'ultimo trimestre del 2013, le previsioni continuano a dare un sensibile aumento delle vendite per questo prodotto, grazie all'ampia fetta di mercato che le SSD stanno a poco a poco sottraendo agli HDD.

Bibliografia

[1] Pavan P., Bez R., Olivo P., Zanoni E., *Flash Memory Cells - An Overview*, 1997, Proceedings of the IEEE

[2] Camerlenghi E., Bez R., Modelli A., Visconti A., *Introduction to Flash Memory*, 2003, Proceedings of the IEEE

[3] Sanvido M. A. A., Chu F. R., Kulkarni A., Selinger R., *NAND Flash Memory and Its Role in Storage Architectures*, 2008, Proceedings of the IEEE

[4] Micheloni R., Picca M., Amato S., Schwalm H., Scheppler M., Commodaro S., *Non-Volatile Memories for Removable Media*, 2009, Proceedings of the IEEE

[5] Saccardi P., *Il futuro instabile delle memorie SSD*, 2013, http://www.tomshw.it/cont/articolo/il-futuro-instabile-delle-memorie-ssd/47286/1.html

[6] Shen J., *NAND market moderates as flash memory loading slackens in devices, says IHS*, 2013, http://www.digitimes.com/news/a20131105PR201.html

[7] Kerekes Z., *Charting the Rise of the Solid Stake Disk Market*, 2013, http://www.storagesearch.com/chartingtheriseofssds.html

[8] Hutchinson L., *The future of flash memory: tiny (and extremely tough to build)*, 2012, http://arstechnica.com/information-technology/2012/07/the-future-of-ssds/