POLITECNICO DI MILANO

V Facoltà di Ingegneria - Ingegneria dell'Informazione

Corso di Laurea Magistrale in Ingegneria Elettronica

Caratterizzazione sperimentale di celle di memoria per array 3D

Relatore: Prof. Alessandro SOTTOCORNOLA SPINELLI

Correlatore: Dott. Elisa CAMOZZI

Tesi di Laurea Magistrale di

Massimo TAGLIAFERRI

matricola 709628

Anno Accademico 2010 – 2011

Indice

Introduzione

Negli ultimi decenni, la richiesta sul mercato di dispositivi elettronici, quali telefoni cellulari, personal computers, lettori MP3, fotocamere digitali, navigatori satellitari, è stata caratterizzata da una crescita esponenziale. Parimenti, si è reso necessario per le aziende un aumento delle attività di ricerca, sviluppo e realizzazione di supporti di memoria veloci e capienti. L'evoluzione tecnologica ha portato alla realizzazione di memorie elettroniche flash, sviluppate con la possibilità di aumentare la capacità di memorizzazione dei dati in aree occupate sempre più piccole attuando i principi dello scaling delle dimensioni e limitando in questo modo costi di produzione. Tuttavia questo trend deve fare i conti con i limiti dovuti alla continua riduzione delle dimensioni delle memorie e delle tensioni con cui vengono alimentate, spingendo la tecnologia verso soluzioni nuove e innovative. Tra le tante possibilità il concetto di memorizzazione dei bit in nodi di storage discreti unito alla possibilità di realizzare celle di memoria "stacked" in tre dimensioni ha attirato una notevole attenzione grazie alla possibilità di aumentare ulteriormente la capacità di memorizzazione utilizzando le tecnologie di memoria di tipo flash.

Il presente lavoro di Tesi di Laurea Magistrale della durata di nove mesi si colloca all'interno della ricerca su celle di memoria floating body come possibile evoluzione delle tradizionali memorie flash e delle memorie di tipo charge trapping, svolta dal Dipartimento di Elettronica e Informazione del Politecnico di Milano in collaborazione con il centro R&D di Numonyx-Micron in Agrate Brianza (MB). In particolare il lavoro è focalizzato sulla caratterizzazione sperimentale delle celle di memoria TANOS (TaN-Allumina-Nitride-Oxide-Silicon) e SONOS (Silicon-Oxide-Nitride-Oxide-Silicon) fabbricate in tecnologia Silicon On Insulator.

Nel primo capitolo verranno presentate le varie tecnologie di memorizzazione non volatile a semiconduttore, focalizzando l'attenzione sulla tecnologia flash tradizionale. Verrà introdotto il transistore a floating gate, mattone fondamentale della tecnologia

flash, e ne verranno descritti la struttura e i principi di funzionamento. Verranno introdotti i concetti di programmazione e cancellazione di una cella, inoltre i concetti di ciclatura e ritenzione legati all'affidabilità di una memoria.

Nel secondo capitolo verranno descritti i principi con i quali viene attuato lo scaling tecnologico, i suoi vantaggi e i suoi limiti. Per superare i limiti che affliggono lo scaling verrà presentata l'idea del charge trapping come possibilità di ingegnerizzazione del concetto di floating gate e il concetto di Silicon On Insulator (SOI), preludio delle memorie stacked e delle memorie 3D.

Nei tre capitoli successivi verranno elencate le misure sperimentali condotte sulle celle di memoria TANOS a substrato contattabile costruite con l'idea del charge trapping, passando per la tecnologia TANOS a substrato flottante fino a giungere alla cella di memoria SONOS a substrato flottante con area attiva arrotondata. In particolare saranno valutate le caratteristiche di programmazione, cancellazione, ciclatura e ritenzione delle singole tecnologie. Verranno messe a confronto le prestazioni di questi dispositivi in relazione ai risultati ottenuti per stabilire se le celle utilizzate possiedono le qualità necessarie per essere utilizzate nello sviluppo di array di memorie 3D.

Lo scopo è quello di comparare le performance dei dispositivi a substrato flottante con le precedenti tecnologie per presentarle come architetture innovative e all'avanguardia nella realizzazione delle memorie elettroniche.

Capitolo 1

Introduzione alle memorie flash

1.1 Memorie a semiconduttore

Negli ultimi decenni il mercato dei dispositivi elettronici ha mostrato una crescita esponenziale: questo ha reso le memorie a semiconduttore un componente indispensabile per tutti i sistemi elettronici moderni. Basti pensare al loro utilizzo nei personal computer, nei sistemi GPS, nelle macchine fotografiche digitali, nei telefoni cellulari e così via, per capire quanto queste memorie siano fondamentali per l'elettronica di largo consumo [1]. Una classificazione delle memorie viene fatta in base alla loro capacità di mantenere l'informazione memorizzata anche se non alimentate. Si distingue quindi tra memorie volatili, come ad esempio le RAM (Random Access Memory), e memorie non volatili, come ad esempio le ROM (Read Only Memory) e le memorie flash. Le memorie non volatili si differenziano tra di loro in base alla flessibilità e alla capacità di poter modificare l'informazione in esse contenuta. Ad esempio, le ROM sono memorie a sola lettura: l'informazione viene immagazzinata (scritta) in fase di fabbricazione e non può essere più modificata in un secondo momento. L'evoluzione delle ROM sono le PROM (Programmable ROM) in cui l'utente può memorizzare l'informazione desiderata senza tuttavia poterla comunque modificare successivamente. Un ulteriore ed importante sviluppo delle memorie ROM sono le EPROM (Erasable Programmable ROM) e le EEPROM o E^2PROM (Elettrically Erasable Programmable ROM) in cui l'informazione viene scritta tramite l'inserimento di cariche in un elettrodo isolato (floating gate) con la possibilità di modificare il dato più volte [2]. Le memorie EPROM possono essere cancellate per effetto fotoelettrico grazie all'esposizione a raggi ultravioletti, mentre le memorie EEPROM vengono

cancellate elettricamente grazie all'effetto tunnel chiamato Fowler-Nordheim. Questo differente meccanismo, che sfrutta un campo elettrico opportunamente applicato per stimolare l'effetto tunnel, è molto più veloce rispetto ai precedenti meccanismi utilizzati nelle memorie ROM; tuttavia le memorie EEPROM hanno lo svantaggio di richiedere una maggiore occupazione di area attiva e maggiori costi di realizzazione (essendo costituite da due transistori).

Le memorie flash sono una pietra miliare dell'evoluzione tecnologica nello sviluppo di memorie non volatili e hanno riscosso enorme successo per le loro caratteristiche di velocità di cancellazione, attuata elettricamente, compattezza e ridotte dimensioni: una cella è costituita da un unico transistor. Tuttavia, i limiti imposti dallo scaling alle memorie flash stanno portando lo sviluppo tecnologico delle memorie verso soluzioni innovative. Una di queste soluzioni vede la sostituzione dell'elettrodo di floating gate con un sottile strato di dielettrico: si parla di memorie MNOS (Metal-Nitride-Oxide-Semiconductor) come per esempio le SONOS e le TANOS in cui il ruolo dell'elettrodo di floating gate è demandato alle trappole presenti nel nitruro. Questo tipo di memorie, dette anche CT (Charge Trapping), permetterebbero in teoria una maggiore scalabilità e potrebbero anche essere utilizzate in strutture tridimensionali a doppio o triplo gate dette FinFlash, perché molto simili ai FinFET se non per l'ossido di silicio sostituito da un dielettrico di tipo ONO (Oxide-Nitride-Oxide).

Esistono anche ricerche rivolte alla realizzazione di dispositivi che sfruttano principi di funzionamento e architetture completamente differenti rispetto alla floating gate, come per esempio le memorie PCM (Phase Change Memory), in cui l'informazione viene immagazzinata per mezzo della variazione dello stato del materiale (amorfo o policristallino) con cui è composto l'elettrodo di controllo (realizzato normalmente in materiali calcogenuri come $Ge_2Sb_2Te_2$), oppure le memorie a nanotubi in carbonio (CNT, Carbon Nano Tubes), in cui l'informazione è legata alla deformazione dei nanotubi in relazione alla diversa tensione applicata.

1.2 La cella a floating gate

Come accennato nel precedente paragrafo, le memorie flash utilizzano un elettrodo ausiliario chiamato floating gate per immagazzinare il dato: la cella di memoria così costruita costituisce il transistore a floating gate, che può essere considerato il mattone fondamentale di queste tecnologie [1]. Il transistore a floating gate è un dispositivo simile al comune transistore MOS, nel quale viene inserito un elettrodo isolato (il floating gate) tra l'elettrodo di gate (chiamato control gate) e il semiconduttore di substrato. Il funzionamento del dispositivo permette, tramite il control gate, di modulare la tensione di soglia V_T del transistore variando la carica contenuta nel floating gate per mezzo di accoppiamenti capacitivi tra il gate di controllo e il gate flottante stesso.

Di seguito verranno descritti brevemente la struttura e il principio di funzionamento della cella a floating gate.

1.2.1 Struttura

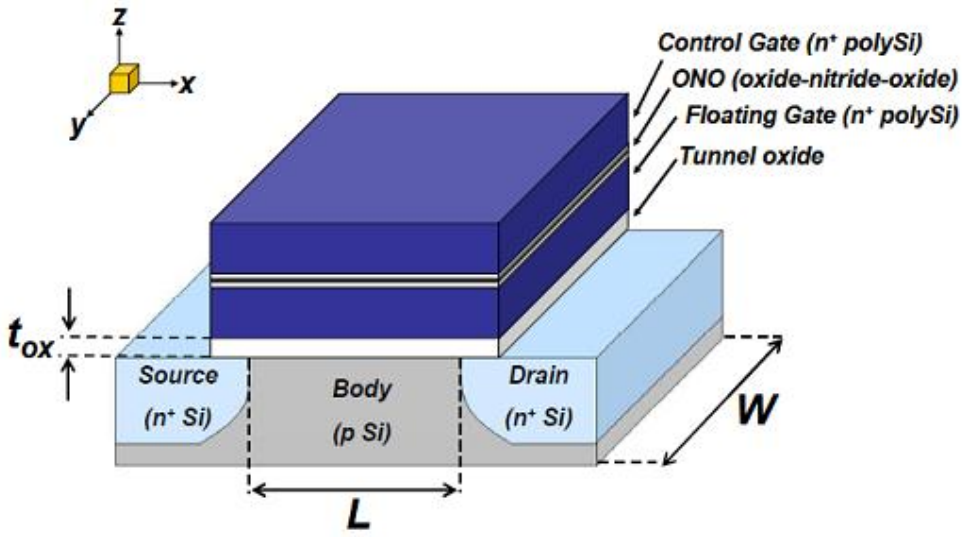


Figura 1.2.1: struttura della cella elementare flash a floating gate

In figura 1.2.1 è riportata la struttura di un transistore a floating gate, cella elementare delle memorie flash. A differenza di un classico transistore MOS, si nota la presenza di uno strato isolato di silicio policristallino n^+ tra il control gate e il substrato: questo floating gate è isolato dal body tramite uno strato di ossido di silicio SiO_2, detto ossido di tunnel, e verso l'elettrodo di comando generalmente tramite un dielettrico di tipo ONO (Oxide-Nitride-Oxide).

1.2.2 Funzionamento

La tensione di soglia V_T della cella a floating gate dipende dalla quantità di carica che si trova nel gate flottante: variando questa carica è possibile modulare la tensione di soglia. Sfruttando questo principio, è possibile legare la memorizzazione del dato alla tensione di soglia della cella attraverso la carica immagazzinata nel floating gate.

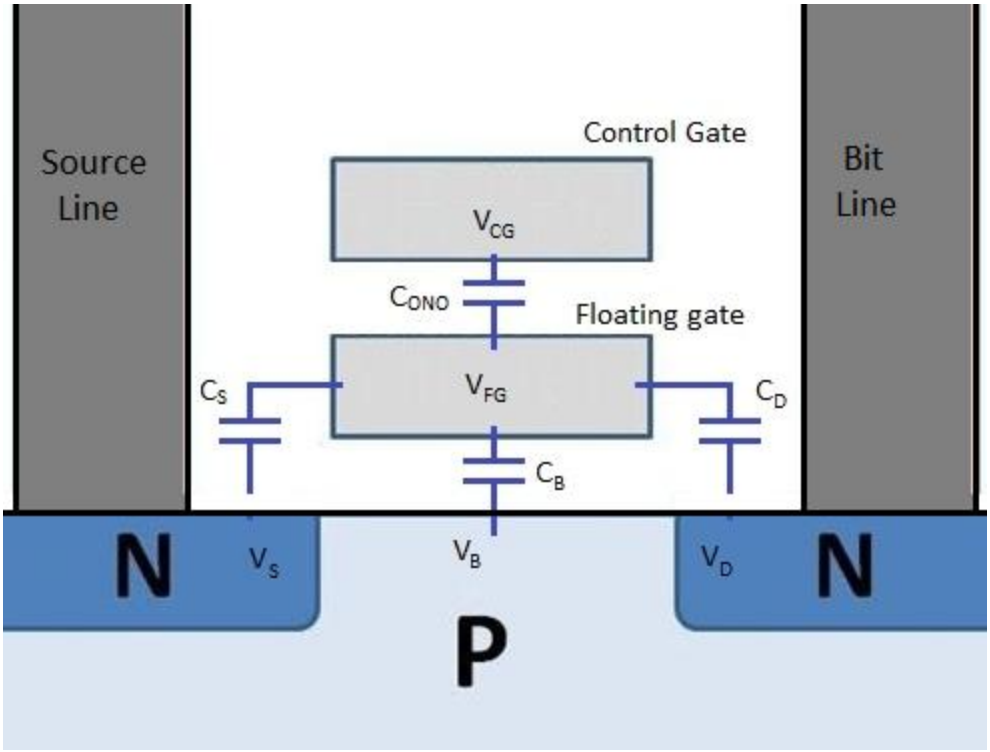


Figura 1.2.2: schema degli accoppiamenti capacitivi nella cella a floating gate

Trascurando in prima approssimazione la capacità parassita associata allo svuotamento del silicio, facendo riferimento alla figura 1.2.2, si può scrivere

$$V_{FG} = \frac{C_{ONO}}{C_{TOT}} V_{CG} + \frac{C_S}{C_{TOT}} V_S + \frac{C_D}{C_{TOT}} V_D + \frac{C_B}{C_{TOT}} V_B + \frac{Q_{FB}}{C_{TOT}} \quad (1.2.1)$$

dove $C_{TOT}=C_{ONO}+C_B+C_S+C_D$ e Q_{FG} è la carica contenuta nel floating gate.
Definiti i coefficienti di accoppiamento capacitivo $\alpha_G = \frac{C_{ONO}}{C_{TOT}}$, $\alpha_B = \frac{C_B}{C_{TOT}}$, $\alpha_S = \frac{C_S}{C_{TOT}}$, $\alpha_G = \frac{C_G}{C_{TOT}}$, si può riscrivere l'espressione della tensione del floating gate come segue:

$$V_{FG} = \alpha_G V_{CG} + \alpha_S V_S + \alpha_D V_D + \alpha_B V_B + \frac{Q_{FB}}{C_{TOT}} \quad (1.2.2)$$

Per legare la carica Q_{FG} alla tensione di soglia riferita al floating gate V_T^{FG} , ovvero alla tensione che permette la formazione del canale di inversione nel silicio al di sotto del gate, bisogna richiamare l'espressione della corrente di canale di un MOS in zona lineare in funzione dei potenziali V_{CG} (tensione della control gate), V_S (tensione dell'elettrodo di source), V_D (tensione dell'elettrodo di drain):

$$I_D = k\left[(V_{CG} - V_T^{CG})V_D - \frac{V_D^2}{2}\right] \quad (1.2.3)$$

e bisogna scrivere l'equazione che lega la tensione di soglia riferita al floating gate, V_T^{FG}, alla tensione di soglia riferita al control gate, V_T^{CG}:

$$V_T^{CG} = \frac{V_T^{FG}}{\alpha_G} - \frac{Q_{FG}}{C_{ONO}} - fV_D \quad (1.2.4)$$

dove $f = \frac{\alpha_D}{\alpha_G}$.
Sostituendo quest'ultima nella (1.2.3) si ricava

$$I_D = k\left\{\left[V_{CG} - \left(\frac{V_T^{FG}}{\alpha_G} - \frac{Q_{FG}}{C_{ONO}}\right)\right]V_D + \left(f - \frac{1}{2\alpha_G}\right)V_D^2\right\} \quad (1.2.5)$$

da cui si ricava la tensione di soglia della cella

$$V_T = \frac{V_T^{FG}}{\alpha_G} - \frac{Q_{FG}}{C_{ONO}} \tag{1.2.6}$$

Il secondo termine della (1.2.6) mostra chiaramente come la tensione di soglia della cella dipenda dalla carica immagazzinata nel floating gate: agendo su tale carica Q_{FG} è possibile modulare la soglia e quindi definire uno stato di immagazzinamento dell'informazione. In particolare, se ad una carica immagazzinata Q_{FG}^L corrisponde una tensione di soglia V_T^L e ad una carica immagazzinata Q_{FG}^H corrisponde una tensione di soglia V_T^H , lo shift della soglia tra i due stati sarà definito come:

$$\Delta V_T = V_T^H - V_T^L = \frac{Q_{FG}^H - Q_{FG}^L}{C_{ONO}} \tag{1.2.7}$$

Avendo uno shift di soglia sufficientemente ampio è possibile riconoscere quindi i due diversi stati della cella floating body: stato programmato e stato cancellato, bit 0 e bit 1.

1.2.3 Programmazione e cancellazione

Per quanto detto nel precedente paragrafo, risulta evidente la tensione di soglia della cella di memoria dipenda dalla quantità di carica immagazzinata nel floating gate. Programmare e cancellare una cella di memoria significa iniettare oppure rimuovere elettroni nel floating gate per modificare lo stato della cella: se ad una certa quantità di carica corrisponde una certa soglia, è possibile quindi distinguere due stati nei quali la cella si può trovare, stato programmato o stato cancellato. Convenzionalmente si associa lo stato "0" alla cella programmata, mentre si associa lo stato "1" alla cella cancellata.

L'operazione di cancellazione, ovvero la rimozione degli elettroni dalla floating gate, sfrutta il fenomeno dell'effetto tunnel Fowler-Nordheim, che si realizza tra l'elettrodo flottante e il substrato attraverso l'ossido di tunnel. L'effetto tunnel Fowler-Nordheim

si basa sul principio per cui non è nulla la probabilità di superare una barriera di potenziale per un elettrone sottoposto ad un alto campo elettrico. Pertanto, un elettrone che si trovi nel floating gate in prossimità dell'interfaccia con l'ossido di tunnel ha la possibilità di superare la barriera dell'ossido e venire rimosso dall'elettrodo flottante. La probabilità di tunnel è legata strettamente alla "trasparenza di barriera", che dipende dalla distribuzione della carica all'interno del floating gate, dallo spessore e dalla forma dell'ossido di tunnel. Per questo, l'ossido viene fatto molto sottile in modo che sia sufficiente applicare un campo elettrico non troppo elevato per permettere all'elettrone di spostarsi nel substrato: in questo modo la carica vede una barriera di potenziale di forma triangolare (come in figura 1.2.3) e non trapezioidale (è questo il caso del tunnel diretto).

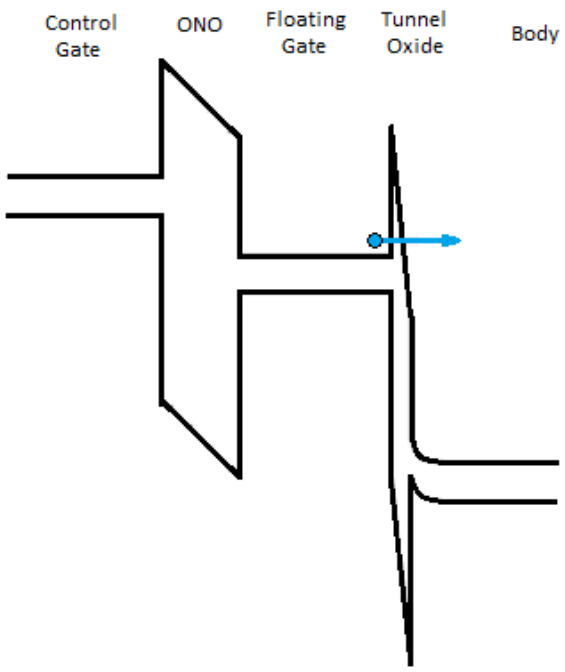


Figura 1.2.3: forma delle bande durante la fase di cancellazione con l'applicazione di un impulso al body

Indicando con F_{OX} $[V/cm]$ il campo elettrico nell'ossido, ϕ_B [V] l'altezza della barriera di potenziale tra il fondo della banda di conduzione nel floating gate e nell'ossido e m_{OX} [kg] la massa efficace degli elettroni nell'ossido (pari a circa $0.5m_0$), si può

scrivere l'espressione della densità di corrente che scorre nell'ossido per effetto tunnel Fowler-Nordheim:

$$J_{F-N} = AF_{OX}^2 \, e^{-\frac{B}{F_{OX}}} \tag{1.2.8}$$

dove

$$A = \frac{q}{16\,\pi^2 \hbar \phi_B}$$

$$B = \frac{4\sqrt{2m_{OX}}}{3\hbar q}(q\phi_B)^{3/2}$$

Al fine di massimizzare l'efficienza del processo di estrazione di cariche dall'elettrodo flottante per tunnel Fowler-Nordheim, l'espressione (1.2.8) mostra come sia opportuno avere spessori di ossido sufficientemente sottili e cadute di tensione sull'ossido elevate, per ottenere un campo elettrico nell'ossido sufficientemente alto da raggiungere la situazione di barriera triangolare.

La programmazione della cella di memoria può essere effettuata sfruttando, come per la cancellazione, l'effetto tunnel Fowler-Nordheim invertendo i segni delle tensioni: in questo caso, con un campo elettrico applicato in senso opposto, sono gli elettroni del substrato vicini all'interfaccia ad avere una probabilità non nulla di attraversare la barriera di potenziale dell'ossido in direzione della floating gate, come in figura 1.2.4

Un altro metodo che permette di programmare la cella, chiamato Channel Hot Electons Injection (CHEI), sfrutta l'iniezione di portatori caldi dal canale verso la floating gate: si basa sull'applicazione di una differenza di potenziale al canale che permette di accelerare gli elettroni al punto da far loro superare la barriera dell'ossido per emissione termoionica nei pressi della zona di drain.

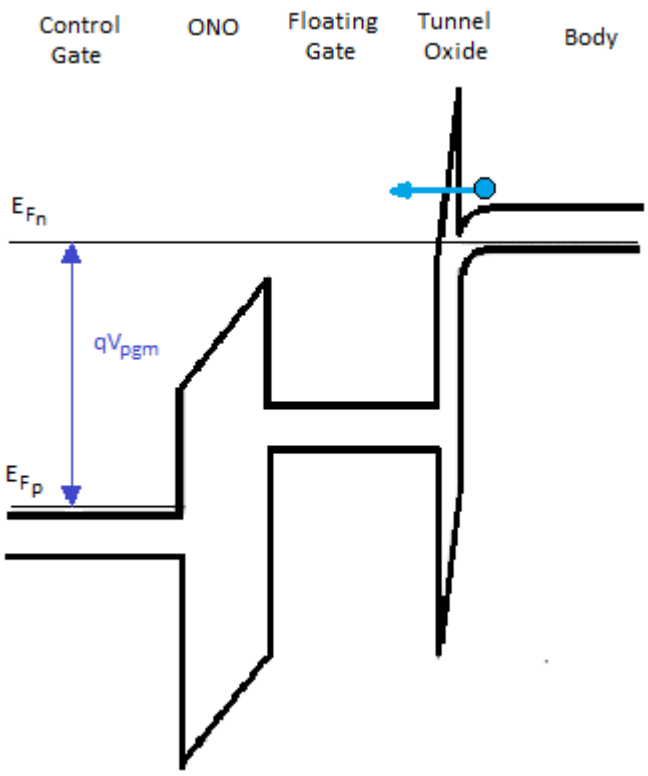


Figura 1.2.4: forma delle bande durante la fase di programmazione per effetto tunnel Fowler-Nordheim

1.2.4 Ciclatura e ritenzione

Ogni ripetizione delle operazioni di programmazione e cancellazione effettuate sulle memorie non volatili causa un degrado della memoria stessa: la caratteristica di ciclatura di una cella di memoria definisce il numero di cicli che il transistor può sostenere senza che il suo danneggiamento sia intollerabile. Una misura di ciclatura consiste nel sottoporre la cella ad una serie di operazioni di programmazione e cancellazione (cicli) andando a monitorare l'andamento della soglia nello stato programmato e nello stato cancellato. Tipicamente, all'aumentare del numero di cicli i valori di soglia degli stato programmato e cancellato si scostano sempre di più dal valore iniziale verso valori via via maggiori. Le prestazioni della memoria vengono quindi influenzate dall'utilizzo ripetuto della memoria stessa, in particolare l'efficienza con cui la memoria viene cancellata peggiora all'aumentare del numero di cicli, poiché a parità di impulso si ottiene un valore di soglia sempre più elevato. Il

degrado più critico di una cella è quello relativo allo stato cancellato, che si avvicina sempre più al livello di lettura fino a superarlo così da non poter più essere discriminato dallo stato programmato.
La capacità di una memoria di mantenere il dato per un lungo periodo di tempo in assenza di alimentazione costituisce una caratteristica fondamentale per l'affidabilità e le prestazioni, e prende il nome di ritenzione. Un dispositivo di memoria ideale sarebbe in grado di mantenere il dato memorizzato (quindi la carica immagazzinata) per un periodo di tempo illimitato. Nel caso reale, la presenza stessa della carica è causa di un leggero campo elettrico e di un debole piegamento delle bande che fanno si che la carica intrappolata possa essere persa col passare del tempo; inoltre, lo spessore dell'ossido di tunnel influenza esponenzialmente la densità di corrente che scorre nell'ossido per effetto tunnel Fowler-Nordheim. Un altro contributo che diminuisce le prestazioni di ritenzione di una memoria è dovuto alla difettosità dell'ossido. Questo fenomeno, che sarà discusso nel dettaglio più avanti, è ancor più evidente nei dispositivi ciclati nei quali il numero di difetti nell'ossido è maggiore perché ogni operazione di cancellazione e programmazione avviene per effetto tunnel proprio attraverso l'ossido di tunnel. Perdere la carica significa perdere l'informazione, perciò si capisce bene come il comportamento di una memoria in ritenzione sia una caratteristica critica delle prestazioni di un dispositivo di immagazzinamento dei dati.

1.3 Architettura delle memorie flash

Nell'introduzione della struttura della cella a floating gate è stato detto che costituisce il mattone fondamentale per la costruzione delle memorie flash. Una memoria è costituita da una matrice di transistori floating gate opportunamente collegati tra loro: le varie architetture differiscono proprio nel modo in cui vengono effettuati questi collegamenti, ottenendo differenti prestazioni e caratteristiche. Al fine di mantenere un maggior controllo legato anche alla gestione e all'affidabilità della memoria e per

semplificare strutture che contengono un elevato numero di transistors, le matrici vengono suddivise in settori o blocchi.

1.3.1 Architettura NOR

Immesse sul mercato per la prima volta nel 1988 da Intel, le memorie flash NOR sono tutt'oggi utilizzate soprattutto in quelle applicazioni che richiedono il salvataggio permanente di dati raramente soggetti a modifiche (ad esempio, sistemi operativi di fotocamere digitali e telefoni cellulari), oppure per contenere i *firmware* dei microcontrollori. Come mostrato in figura 1.3.1, in questa architettura la Bit Line collega in parallelo tutti i drain delle celle sulla stessa colonna, le Word Lines collegano tutti i control gate dei transistori sulla stessa riga e le Source Lines collegano tutti i terminali di source.

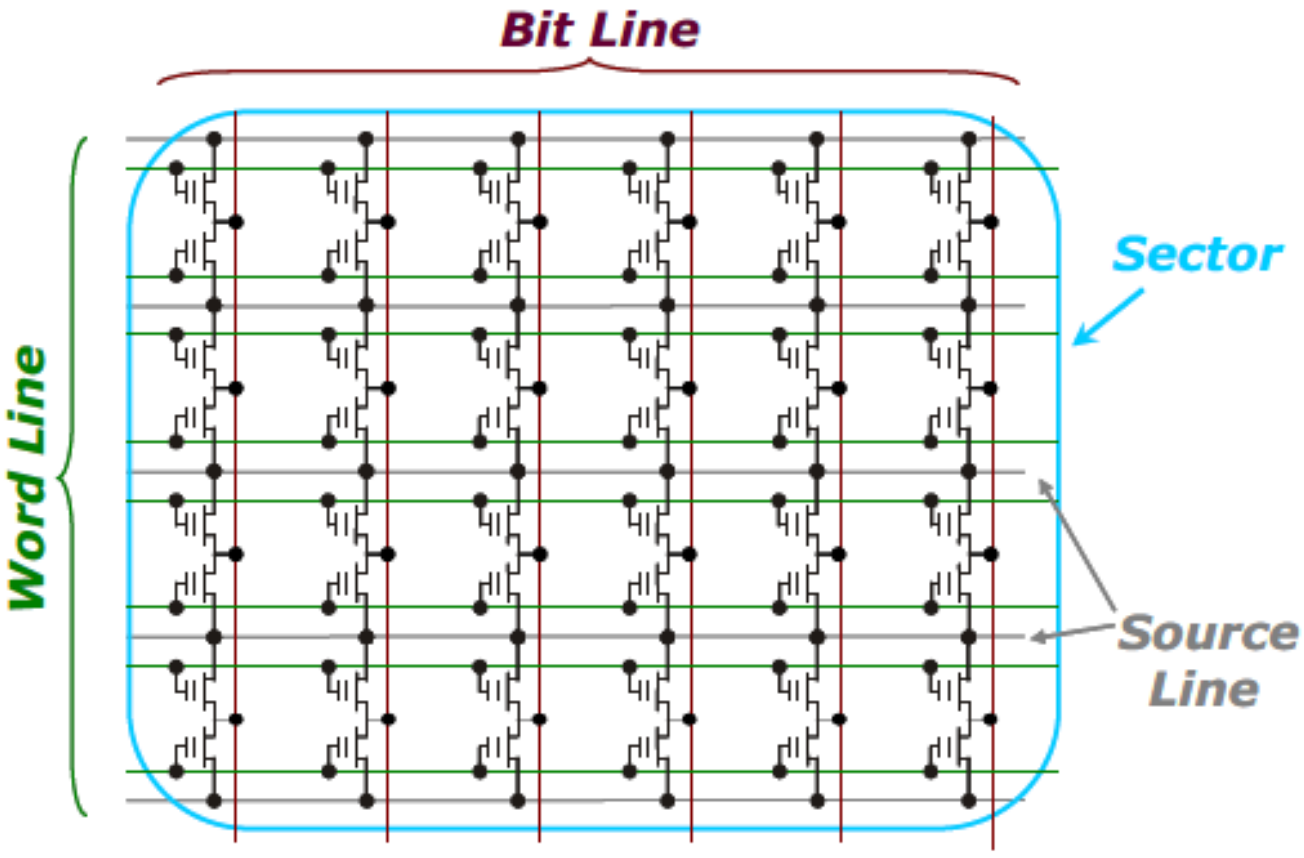


Figura 1.3.1: architettura di un settore di una memoria flash NOR

In questo tipo di architettura ogni settore è delimitato da opportune diffusioni n^+ che opportunamente polarizzate possono dividere e isolare i vari blocchi della matrice: questo tipo di impianto è chiamato *triple well* (fig. 1.3.2).

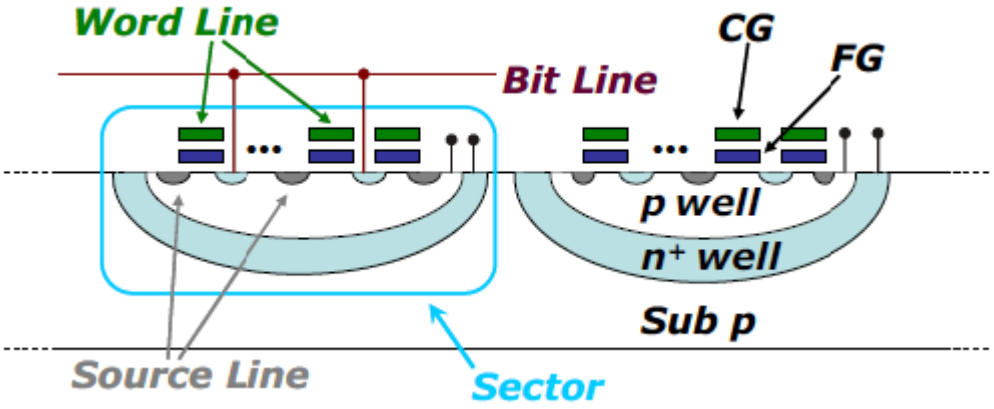


Figura 1.3.2: schema di un impianto *triple well*

Le celle di memoria flash NOR vengono programmate singolarmente sfruttando l'effetto di iniezione di elettroni caldi (CHEI): viene polarizzata alla tensione V_{pgm} la WL della cella da programmare e alla tensione V_{DS} la corrispondente BL, lasciando a massa le altre WLs e BLs, la SL e il body. In questo modo solo la cella selezionata viene programmata: per quanto riguarda le altre celle che stanno sulla stessa colonna, avendo polarizzato la control gate a massa, il campo elettrico dovuto alla differenza di potenziale ai capi dell'ossido di tunnel non favorisce il passaggio di elettroni verso la loro floating gate; le altre celle che stanno sulla stessa riga hanno la BL polarizzata a massa e quindi all'interfaccia non ci sono elettroni energetici nel canale che possano passare all'elettrodo flottante.

La cancellazione delle memorie flash NOR avviene per effetto tunnel Fowler-Nordheim: le BL vengono lasciate flottanti, viene applicata una tensione negativa alle WL e una tensione positiva al body, cortocircuitato con le SL. A differenza di quanto avviene con la programmazione che riguarda una singola cella, poiché le celle di un blocco condividono il substrato tra loro, la cancellazione riguarda un intero settore (gli altri settori invece non risentono dell'effetto di cancellazione grazie all'isolamento

triple well). Lo stato cancellato di una memoria NOR sarà sempre a tensioni positive per questioni legate all'architettura e al processo di lettura. Per questo dopo ogni cancellazione si ha una fase detta *soft-programmazione* per riportare selettivamente le soglie di tutte le celle del settore cancellato a valori positivi.

L'operazione di lettura delle memorie flash NOR è semplice e veloce. Si applica una opportuna tensione V_{read} alla WL della cella che si vuole andare a leggere, SL e body vengono portati a massa e la BL è connessa al sistema di sensing. Nel caso di cella nello stato cancellato, si applica $V_{read} > V_T$ (con V_T tensione di soglia della cella) perciò si formerà il canale sotto il gate del transistor permettendo alla corrente di scorrere verso la BL; nel caso di cella programmata non si forma il canale e non scorre corrente perché il transistor è spento. Le celle che stanno sulla stessa WL della cella che viene letta sono afferenti alle altre BL opportunamente polarizzate a massa, perciò non intervengono nell'operazione di lettura; i transistors collegati alla cella che viene letta con la stessa BL non intervengono perché sono stati in precedenza soggetti all'operazione di *soft-programmazione* e avendo polarizzato le WL a massa essi sono sicuramente spenti, senza interferire nell'operazione di lettura. Inoltre, sfruttando più sistemi di sensing, è possibile effettuare letture in parallelo di più celle, possibilità che diminuisce sensibilmente il tempo richiesto per la lettura della memoria.

La semplicità e la velocità con cui l'operazione di lettura può essere condotta costituisce uno dei principali vantaggi delle architetture flash NOR; d'altro canto, il principale difetto di questo tipo di architettura sta nella necessità di avere un contatto di drain per ogni due transistors affacciati e nel dover conseguentemente collegare tutti i drain in parallelo, caratteristica che aumenta l'area occupata da ogni singola cella e quindi l'ingombro della memoria stessa.

1.3.2 Architetture NAND

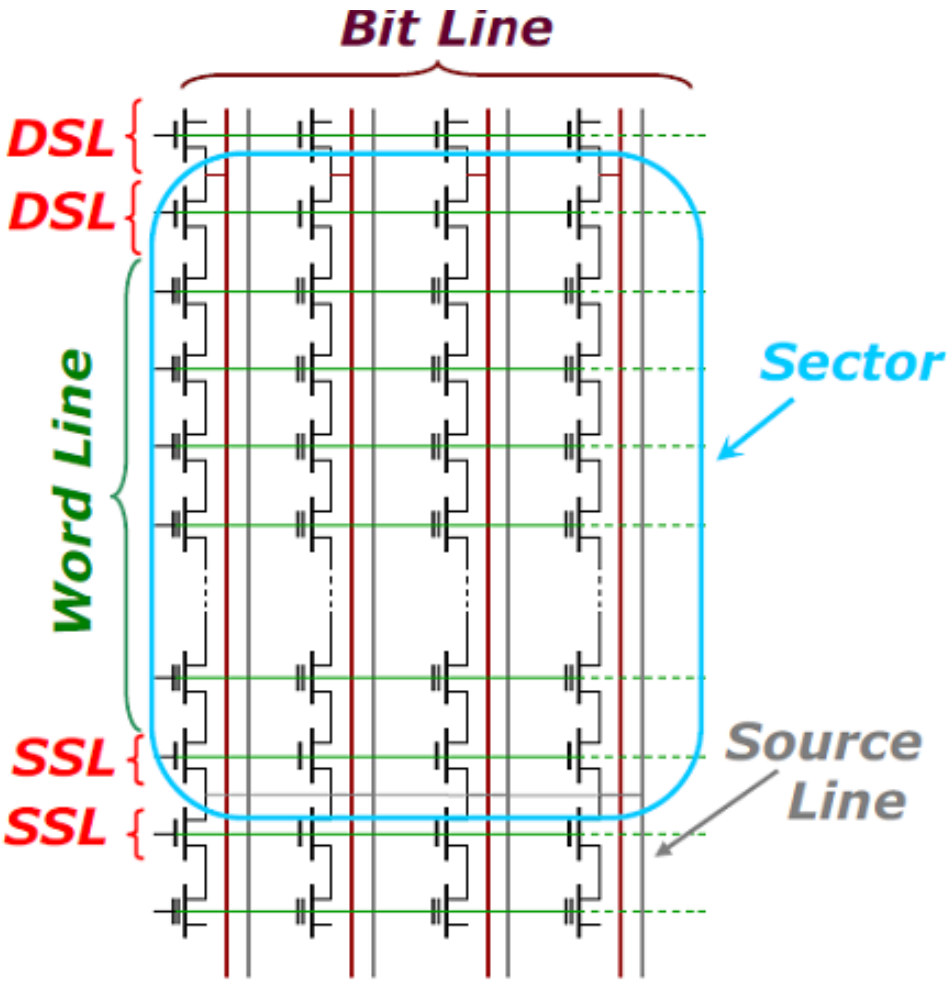


Figura 1.3.4: architettura di un settore di una memoria flash NAND

Inventate nel 1989, le memorie flash NAND si sono imposte nel mercato sulle NOR per le loro caratteristiche di velocità, prestazioni elevate e costi ridotti. L'architettura di connessione delle celle nelle memorie NAND permette di superare il problema di grande occupazione di area delle NOR permettendo di integrare un maggior numero di transistor per unità di area in grado di immagazzinare i dati; inoltre, un grande vantaggio è dato dal fatto che, nonostante i tempi per le operazioni di programmazione e cancellazione sulla singola cella siano più lunghi rispetto alle NOR, il tempo totale di programmazione per una memoria NAND sia notevolmente minore.

E' infatti possibile programmare più transistors in parallelo senza consumo di potenza statica, ottenendo un *throughput* (quantità di dati trasmessi nell'unità di tempo) maggiore e prestazioni di gran lunga più elevate rispetto a quelle ottenibili con le NOR. Queste caratteristiche hanno fatto diventare le memorie NAND il principale supporto di *data storage* per le applicazioni multimediali di largo consumo.

In figura 1.3.4 si può notare una tipica architettura di connessione di transistors costituente una memoria NAND. Le celle di una stessa colonna sono collegate in serie (stringa) e ogni serie di transistors è collegata alla propria Bit Line tramite un transistor di selezione DSL (Drain Selector Line); i gate di controllo delle celle della stessa riga sono collegati in parallelo tramite le Word Lines. I settori o blocchi di memoria sono separati tra loro grazie ai transistori di selezione DSL e SSL (Source Selector Line) che, quando vengono attivati, consentono il collegamento delle varie colonne alle BL (attraverso DSL) e alla SL (source line, attraverso SSL); nel caso in cui non siano attivi, il settore resta isolato dal resto della memoria.

Ne risulta una struttura semplice e compatta, un numero piuttosto ridotto di contatti, inoltre la connessione in serie tra due transistors adiacenti sulla stessa colonna permette loro di condividere la regione n^+, riducendo ulteriormente l'area occupata dalla memoria.

L'operazione di programmazione di una memoria flash NAND sfrutta l'effetto tunnel Fowler-Nordheim degli elettroni di canale verso la floating gate. Poiché durante la fase di programmazione non si ha flusso di corrente nel canale, il notevole vantaggio che ne deriva è la possibilità di programmare con potenza statica dissipata pressoché nulla; tuttavia occorrono elevate tensioni ed elevati tempi di programmazione per attivare l'effetto tunnel Fowler-Nordheim sulla singola cella, per questo sono necessarie soluzioni a livello circuitale (ad esempio possono essere utilizzate delle pompe di carica per ottenere tensioni elevate e si possono programmare più celle contemporaneamente per minimizzare i tempi di lavoro).

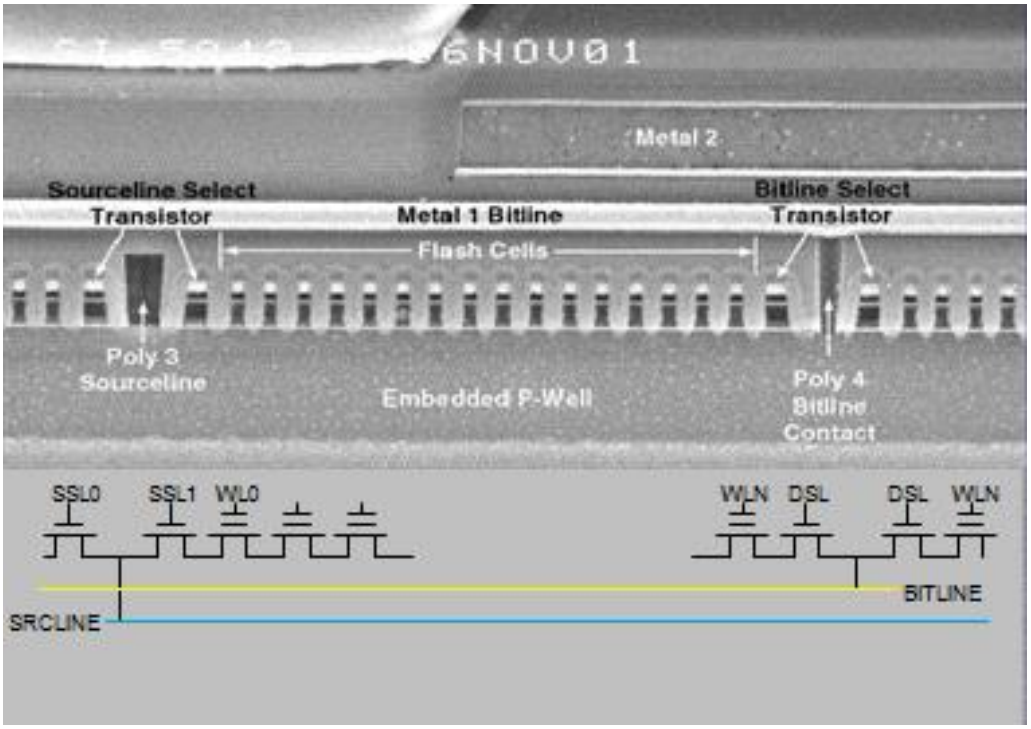


Figura 1.3.5: sezione di una stringa di memoria NAND (tecnologia 120nm) al microscopio elettronico.

La programmazione di una memoria NAND viene effettuata applicando una tensione V_{pgm} alla WL selezionata e mantenendo a massa la corrispondente BL; in questa fase il DSL è attivo in modo da portare a massa la BL mentre il SSL è disattivato lasciando flottanti le colonne: in questo modo non vi è flusso di corrente statica lungo le stringhe. Le rimanenti WL vengono portate ad una tensione V_{pass} allo scopo di aumentare il potenziale del canale delle celle da non programmare (principio del *self-boosting* [3] per effetti di partizione capacitiva tra control gate, floating gate e canale): in questo modo la caduta di tensione effettiva ai capi della barriera ossido-substrato si riduce e le altre celle lungo la stessa WL (portata a V_{pgm}) non vengono programmate. Per aumentare questo effetto vengono anche portate le BL delle celle non interessate alla tensione di alimentazione. Le celle appartenenti alla stessa BL della cella da programmare vengono a loro volta portate a V_{pass}. Il valore ottimale per V_{pass} deve essere tale per cui le celle sulla stessa BL non si programmino (si vorrebbe V_{pass} bassa) e allo stesso modo si riesca ad inibire la programmazione delle celle sulla stessa WL (si vorrebbe V_{pass} alta).

Anche l'operazione di cancellazione delle NAND sfrutta l'effetto tunne Fowler-Nordheim e, come per le NOR, la cancellazione riguarda tutto il settore e non la singola cella. Durante la fase di cancellazione, il body è polarizzato alla tensione V_{erase}, le WL del settore da cancellare vengono portate a massa mentre tutti gli altri contatti vengono lasciati flottanti: in questo modo la cancellazione degli altri settori è inibita perché, per accoppiamento capacitivo con le WL flottanti, non c'è flusso di elettroni dal floating gate verso il substrato grazie alla riduzione della caduta di tensione ai capi dell'ossido.

Lo stato cancellato di una memoria NAND, a differenza di quanto detto in precedenza per le NOR, può aggiungere valori di tensione di soglia negativi, con conseguente allargamento della finestra di lavoro e influenza sull'operazione di lettura della cella.

In fase di lettura, la memoria NAND viene polarizzata come segue: la BL di interesse viene precaricata ad una tensione positiva (ad esempio 1V) in una prima fase in cui SSL e DSL sono attivi e SL viene posta a massa, le altre WL vengono portate a una tensione V_{pass}tale per cui i transistors che si trovano sulla stessa BL di quello di interesse siano accesi (indipendentemente dal fatto che siano programmati o cancellati) così da non interferire nell'operazione di lettura, alla WL di interesse viene applicata la tensione V_{read}.

In questo modo si è in grado di discriminare lo stato della cella: se si trova nello stato cancellato, il transistor sarà accesso perché $V_{CG} = V_{read} > V_T$ e il suo canale sarà in inversione, permettendo lo scorrere di una corrente tra BL e SL; tramite un opportuno sistema di sensing si può rilevare questa corrente e assegnare alla cella il valore logico "1". Viceversa, se la cella si trova nello stato programmato, il transistor sarà spento perché $V_{CG} = V_{read} < V_T$, non ci sarà nessun canale di inversione e non scorrerà nessuna corrente nella stringa; il sistema di sensing non misurerà nessuna corrente e assegnerà alla cella il valore logico "0".

Come per la programmazione, nelle memorie flash NAND è possibile effettuare l'operazione di lettura in parallelo sulle celle di una stessa riga appartenenti ad uno stesso settore.

1.4 Conclusioni

In questo capitolo sono stati introdotti i principali dispositivi di memoria a semiconduttore, focalizzando l'attenzione su dispositivi di memorizzazione non volatile. Dopo una presentazione del tutto generale sui vari dispositivi di memoria, ci si è soffermati sulla tecnologia dominante sul mercato, ovvero la tecnologia flash a floating gate. Dopo aver spiegato i principi di funzionamento di una memoria a floating gate e i principi su cui si basano le operazioni di programmazione e cancellazione, sono state introdotte le due architetture principali per la realizzazione pratica dei dispositivi di memoria (NOR e NAND) confrontando pregi e difetti di entrambe.

Nel prossimo capitolo vedremo come lo scaling tecnologico consenta di realizzare dispositivi sempre più piccoli e permetta di raggiungere i nodi tecnologici predetti dalla legge di Moore, tuttavia questo trend è costretto ad affrontare diversi problemi per i quali verranno presentate alcune soluzioni, come le celle che sfruttano l'idea del Charge Trapping e i transistori Silicon On Insulator.

Capitolo 2
Scaling, charge trapping e post floating gate

2.1 Scaling

L'esigenza di riuscire ad integrare un elevato numero di dispositivi su aree di silicio sempre maggiori ha reso necessario lo sviluppo di tecnologie e modelli in grado di ridurre le dimensioni dei transistors e al contempo ottimizzare le loro prestazioni. L'evoluzione della tecnologia CMOS ha seguito il trend dello scaling con il fine di migliorare la velocità delle operazioni, la minimizzazione della potenza dissipata e il numero di celle di memoria integrabili per unità di area. Questi miglioramenti si accompagnano tuttavia ad alcuni svantaggi tra cui gli effetti di canale corto, l'effetto DIBL, la saturazione della velocità e la modulazione della lunghezza di canale. E' necessario quindi sviluppare nuovi design per i dispositivi, tecnologie e strutture per controllare e superare questi limiti e sfruttare tutti i miglioramenti che le nuove tecniche litografiche sono in grado di apportare. [4]

2.1.1 Principi generali

Grazie ai miglioramenti tecnologici degli ultimi decenni nella fabbricazione dei wafer, è stato possibile realizzare dispositivi che seguissero i principi dello scaling. Fino ad oggi la riduzione delle dimensioni dei dispositivi dovuta ai miglioramenti tecnologici è stata ben modellizzata dalla legge di Moore, secondo la quale il numero di transistors realizzabili su un circuito integrato raddoppia circa ogni due anni.

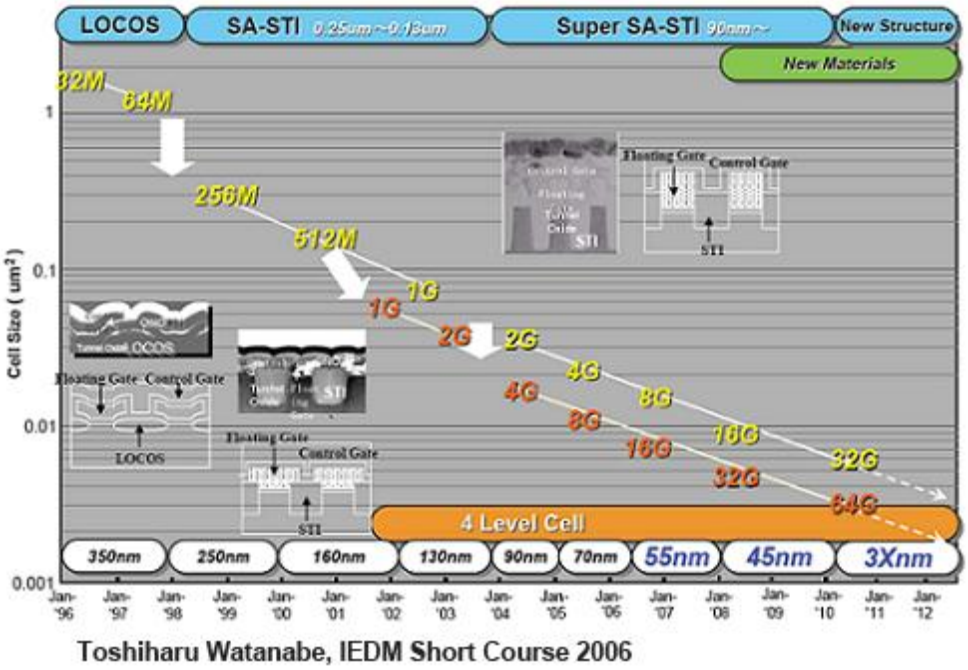


Figura 2.1.1: trend di scaling per le memorie flash secondo la legge di Moore [5]

Utilizzando transistors scalati, le cui dimensioni vengono via via ridotte, è possibile aumentare la densità di celle realizzabili con un conseguente aumento di capacità di memorizzazione dati del dispositivo. Analogamente le prestazioni delle memorie relative alla velocità di processo, al consumo di potenza e alla quantità di dati immagazzinabili ricevono i benefici di questo trend tecnologico: un maggior numero di celle significa un maggior numero di dati immagazzinabili. La progressiva diminuzione negli anni delle dimensioni dei transistors ha permesso il raggiungimento di alcuni nodi tecnologici attorno ai quali si è sviluppata la ricerca per la realizzazione dei dispositivi elettronici: per esempio, i nodi tecnologici di 90nm, 45 nm e 32 nm sono ormai una realtà attuale, la ricerca si sta spingendo verso dimensioni di *half-pitch* che arrivano ai 22 nm ed oltre.

2.1.2 Limiti e problemi

Nonostante gli innegabili benefici dello scaling, i transistors submicrometrici e nanometrici hanno manifestato alcuni limiti dovuti alle loro ridotte dimensioni. In

generale, l'aumento della corrente di off del transistor, dovuta all'aumento degli effetti di canale corto (Short Channel Effects) per effetto DIBL e degrado della pendenza sottosoglia, rappresenta un limite significativo per lunghezze effettive di canale più corte di 15nm. La riduzione dello spessore dell'ossido per migliorare il controllo del gate sul canale si paga in un aumento della corrente di perdita di gate; l'aumento dei drogaggi necessari comporta un degrado della mobilità del canale dovuto allo scattering. Inoltre, ridurre il pitch del gate aumenta i contributi delle capacità parassite contatto-gate e gate-strato epitassiale, incrementando la capacità gate-source del dispositivo. Per quanto riguarda le prestazioni delle celle di memoria, riducendo le dimensioni gli aspetti maggiormente critici riguardano la ritenzione e l'accoppiamento laterale.

Come accennato nel precedente capitolo, un elemento da ricordare nella valutazione delle prestazioni di ritenzione di una cella di memoria è lo spessore dell'ossido di tunnel. Infatti, riducendo le dimensioni, la probabilità di effetto tunnel per un elettrone contenuto nel floating gate aumenta, con conseguente possibile perdita del dato memorizzato.

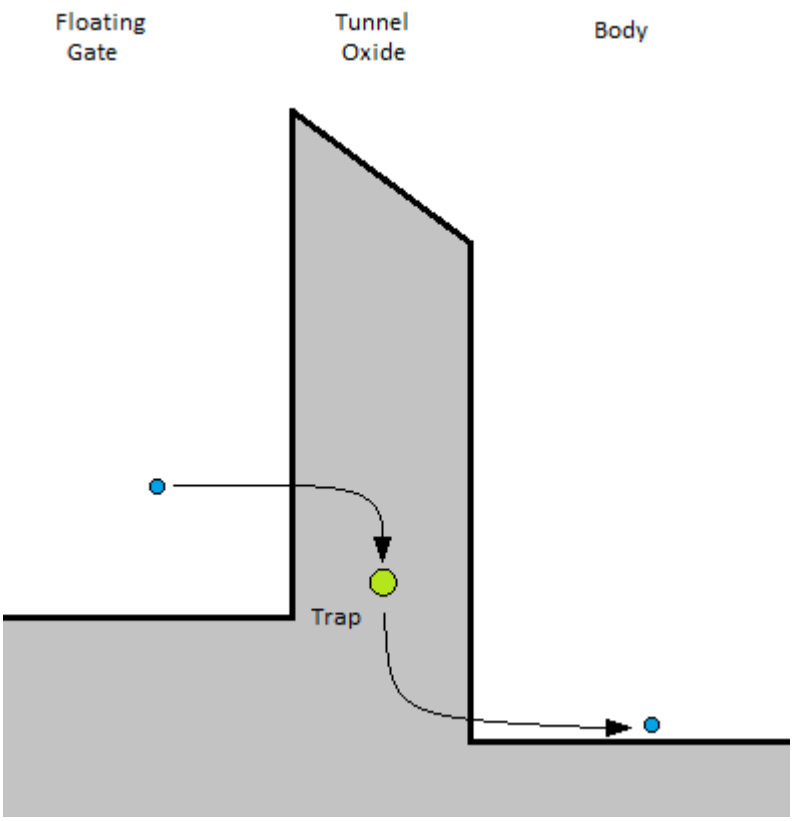


Figura 2.1.2: schematizzazione del Trap Assisted Tunneling

Questo tipo di degrado non è l'unico che contribuisce alla diminuzione delle prestazioni di ritenzione della cella di memoria: un altro aspetto fondamentale è la qualità dell'ossido di tunnel, le cui imperfezioni sono causa di una corrente di perdita non desiderata (*Stress Induce Leakage Current*, SILC). La presenza di un difetto all'interno dell'ossido può fungere per la carica da stato intermedio per mezzo del quale un elettrone potrebbe attraversare la barriera di potenziale sfruttando in una prima fase lo stato energetico nell'ossido (l'imperfezione) e raggiungendo il substrato in una seconda fase. Se non vi fossero imperfezioni nell'ossido, la carica potrebbe attraversare la barriera di potenziale solamente per effetto tunnel diretto; in presenza di difetti, la corrente di perdita aumenta poiché vi è un altro contributo che va a sommarsi a quello di attraversamento diretto della barriera: si parla in questo caso di *Trap Assisted Tunneling*, schematizzato in figura 2.1.2.

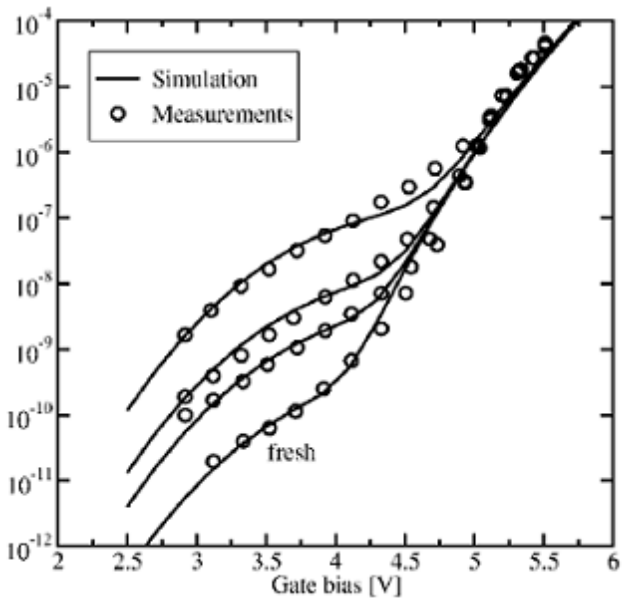


Figura 2.1.3: correnti di leakage al variare del numero di trappole presenti nell'ossido. A bassi campi si vede il contributo dominante del TAT [6]

Al diminuire dello spessore dell'ossido, la corrente di perdita aumenta in maniera esponenziale, ma a bassi campi elettrici si osserva il fondamentale contributo dovuto alla difettosità (figura 2.1.3): tutto questo contribuisce a ridurre le prestazioni di ritenzione della cella.

Un'altra problematica relativa allo scaling è quella dell'accoppiamento laterale tra celle vicine nello stesso settore, chiamato *crosstalk* e schematizzato in figura 2.1.4. Durante un'operazione di programmazione, la WL d'interesse viene portata ad un potenziale elevato e, come evidenziato nel primo capitolo, per accoppiamento capacitivo il floating gate della cella programmata si porta anch'esso ad un potenziale elevato. Poiché le dimensioni della stringa sono ridotte, si ha un effetto spurio ovvero disturbo per accoppiamento capacitivo sui floating gate delle celle appartenenti a WL e BL adiacenti tale per cui un elettrodo flottante che non deve essere programmato si porta in uno stato per cui è favorita l'iniezione di carica. Ne consegue una programmazione non desiderata delle celle adiacenti a quella di interesse.

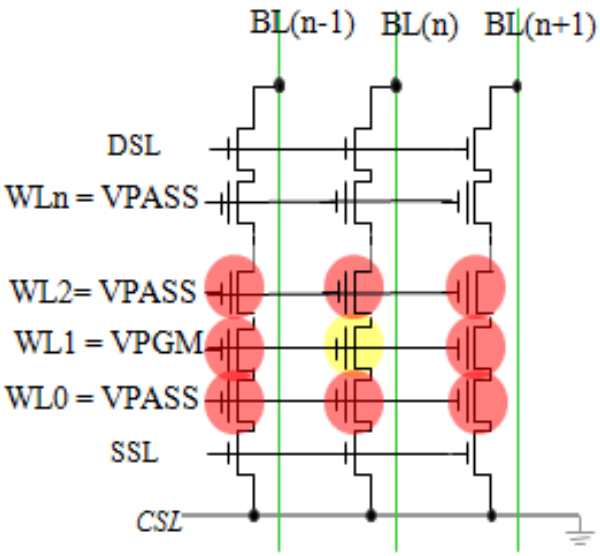


Figura 2.1.4: schema dell'accoppiamento laterale tra celle vicine: evidenziate in rosso le celle affette da crosstalk

2.2 Charge Trapping

Come visto nel precedente paragrafo, all'aumentare dello scaling le prestazioni delle celle di memoria migliorano ma emergono anche i problemi intrinseci alla riduzione delle dimensioni. Secondo *l'International Technology Roadmap of Semiconductors* le memorie Flash oltre i 25nm hanno diversi problemi di integrazione, dovuti principalmente alla riduzione delle dimensioni. Assottigliare l'ossido implica un aumento del rischio di perdita dei dati per l'aumento delle correnti di leakage attraverso il gate; la diminuzione della distanza tra celle adiacenti nell'array causa crosstalk, dovuto alla carica immagazzinata nel floating gate; l'aumento della corrente di off, a causa dell'effetto DIBL (Drain-Induced Barrier Lowering) e del degrado della pendenza di sottosoglia enfatizzati dagli effetti di canale corto (Short Channel Effects), rappresenta un limite significativo per lunghezze effettive di canale inferiori ai 15nm [7]. Per cercare di risolvere questi problemi esistono due principali linee guida: proporre nuove strutture e architetture con innovativi concetti di funzionamento oppure introdurre nuovi materiali; queste linee guida sono sviluppate da tutte le più grandi industrie produttrici di memorie in tutto il mondo anche in maniera combinata. In questo contesto si inserisce l'idea del charge trapping, ovvero la possibilità di

sostituire al floating gate in silicio policristallino uno strato per immagazzinare le cariche tramite trappole discrete, come ad esempio uno strato di nitruro di silicio [8]. Si ottiene in questo modo una cella a trappole discrete in cui la carica viene immagazzinata in nodi di storage discreti indipendenti e isolati, non più in una struttura "continua" come il polisilicio: si riescono quindi a ridurre i contributi delle imperfezioni nell'ossido di tunnel alla corrente di perdita, poiché hanno effetto solamente sulle trappole in prossimità della superficie e lasciano inalterate le altre che mantengono la loro carica. Conseguentemente gli effetti indesiderati relativi alle perdite per SILC sulla tensione di soglia della cella si riducono poiché solo una piccola frazione della carica viene persa per TAT. Viene in questo modo migliorata l'affidabilità della ritenzione in modo da poter ridurre lo spessore dell'ossido e conseguentemente tutte le altre dimensioni della cella seguendo i principi dello scaling.

Esistono diverse tipologie di celle a trappole discrete in dipendenza dalla composizione dello stack del gate, per esempio le celle SONOS (Silicon Oxide Nitride Oxide Silicon) e le celle TANOS (TaN Alumina Nitride Oxide Silicon). Il principio di funzionamento è del tutto analogo a quello della cella a floating gate: la tensione di soglia che influenza la corrente di canale viene modulata dalla presenza di carica all'interno dello strato di immagazzinamento, che nel caso del charge trapping è costituito dal nitruro di silicio. Nel caso di questo tipo di celle, essendo presenti due strati di ossido ben differenti, si fa riferimento all'ossido di top (top oxide) quando si parla dello strato tra il gate di controllo e il nitruro, mentre per ossido di bottom (bottom oxide) si intende lo strato tra il nitruro e il substrato in silicio.

2.2.1 Celle di memoria SONOS

In figura 2.2.1 è mostrata la composizione dello stack della cella di memoria SONOS mentre in figura 2.2.2 è mostrato il diagramma delle bande in condizione di flat band: come si può notare, la struttura è analoga alla classica cella a floating gate con la differenza dello strato di immagazzinamento composto da nitruro di silicio invece del silicio policristallino.

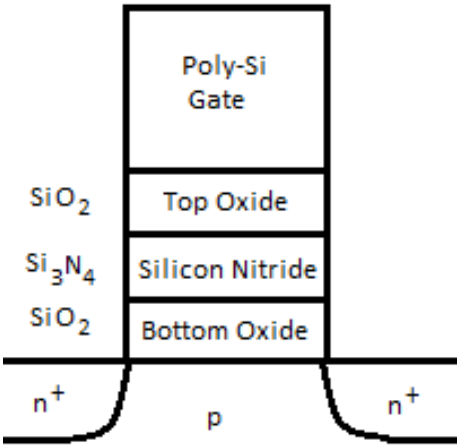


Figura 2.2.1: composizione dello stack di una cella di memoria SONOS

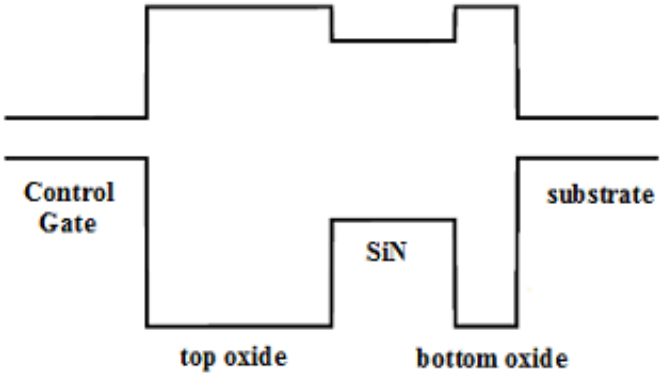


Figura 2.2.2: diagramma delle bande di una cella SONOS al flat band

Dal punto di vista operativo, i principi che governano il funzionamento di questo tipo di cella sono del tutto analoghi a quelli della cella a floating gate: la programmazione avviene per iniezione di carica dal substrato per effetto tunnel Fowler-Nordheim attraverso l'ossido di bottom intrappolando la carica nello strato di nitruro per mezzo di una corretta polarizzazione; la cancellazione avviene sempre per effetto tunnel Fowler-Nordheim dal nitruro al substrato estraendo la carica dalle trappole tramite l'applicazione di un opportuno campo elettrico.

I miglioramenti delle celle SONOS riguardano la possibilità di ridurre ulteriormente le dimensioni e le tensioni applicabili grazie alla presenza di trappole discrete nel nitruro; inoltre, poiché lo strato di immagazzinamento è un dielettrico, spariscono tutti

gli effetti dovuti ad accoppiamenti capacitivi indesiderati tra celle adiacenti come l'accoppiamento laterale (*crosstalk*).

Il principale difetto di questa tecnologia si presenta in fase di cancellazione e prende il nome di *erase saturation*: a causa del sottile ossido di top, si ha un'iniezione di carica tra il control gate e lo strato di nitruro. Questa iniezione può essere tale da bilanciare l'effetto di cancellazione che estrae la carica dal nitruro e la porta nel substrato, imponendo un limite sulla minima quantità di carica che si trova nello strato di intrappolamento, condizione che si riverbera sulla minima tensione di soglia della cella: l'efficacia della cancellazione viene quindi limitata e satura al valore minimo che la tensione di soglia della cella può assumere. Per superare questo problema si è cercato di ridurre lo spessore dell'ossido di bottom in modo da favorire l'iniezione di lacune dal substrato al nitruro durante la cancellazione: in tal modo le cariche positive provenienti dal substrato e presenti nel nitruro possono ricombinarsi con quelle negative iniettate dal control gate, bilanciando questo effetto spurio al fine di aumentare la differenza tra la soglia programmata e la soglia cancellata. Tuttavia, un'eccessiva riduzione dell'ossido di bottom causa una diminuzione delle prestazioni di affidabilità poiché la perdita di carica in ritenzione si fa più rapida, quindi è necessario stabilire uno spessore dell'ossido di bottom opportuno per bilanciare questi due effetti.

2.2.2 Celle di memoria TANOS

Per superare la problematica dell'*erase saturation* analizzata alla fine del precedente paragrafo si è pensato di introdurre materiali alternativi con lo scopo di limitare la corrente di iniezione di carica dal gate al nitruro in fase di cancellazione, cercando di non peggiorare le caratteristiche di ritenzione della cella. Le celle di memoria TANOS sono analoghe alle SONOS in cui però vengono cambiati gli spessori e i matriali che compongono lo stack.

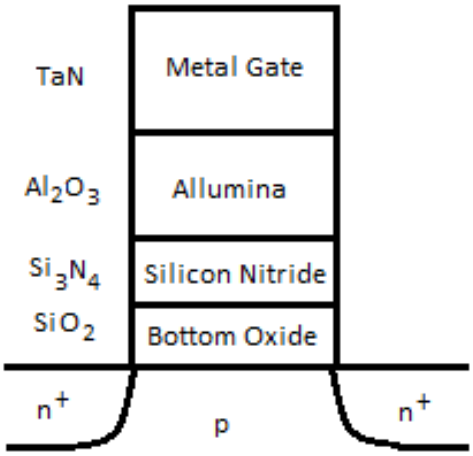


Figura 2.2.3: composizione dello stack di una cella di memoria TANOS

Le celle TANOS utilizzano come ossido di top un materiale high-k, ovvero con elevata costante dielettrica: in questo modo è possibile aumentare lo spessore dell'ossido con l'effetto, a parità di tensione di gate, di diminuire il campo elettrico che causa l'iniezione di carica tra gate e nitruro mantenendo invariata la capacità (poiché $C = {\varepsilon_{OX}}/{t_{OX}} = {\varepsilon_{hk}}/{t_{hk}}$). Il materiale utilizzato è l'ossido di alluminio o allumina (Al_2O_3) per la sua facilità di integrazione con il processo CMOS. Inoltre, per diminuire l'iniezione di elettroni dal gate al nitruro si cerca di alzare la barriera di potenziale vista dagli elettroni realizzando un gate metallico avente un'elevata funzione lavoro e facilmente integrabile con il processo CMOS, ovvero il nitruro di tantalio (TaN). In questo modo viene aumentata l'altezza della barriera di potenziale vista dagli elettroni nel gate e conseguentemente viene diminuita la corrente di tunnel attraverso l'ossido di top.

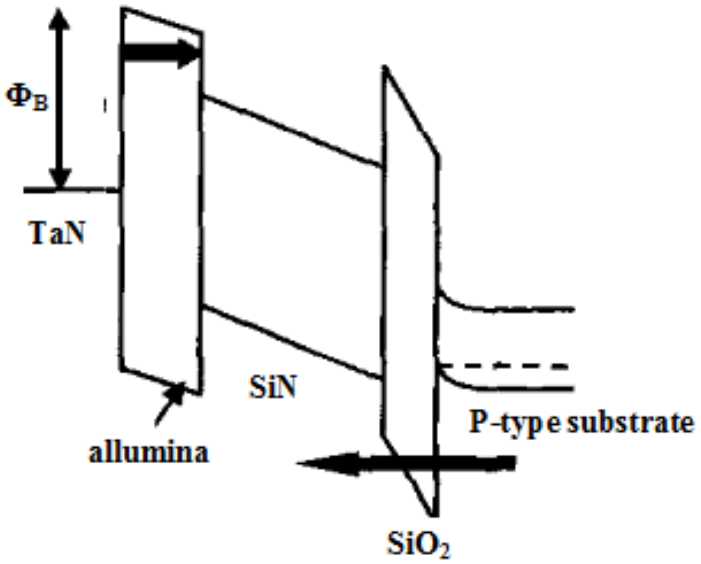


Figura 2.2.4: diagramma delle bande di una cella TANOS in fase di cancellazione

Un ulteriore vantaggio che si ottiene introducendo un gate metallico come gate di controllo è relativo alla programmazione: a parità di tensione di comando applicata, le celle con gate metallico hanno uno scarto di soglia maggiore rispetto alle celle con gate in polisilicio. Questo effetto è dovuto al fatto il metallo utilizzato ha una funzione lavoro maggiore rispetto al polisilicio pertanto la tensione di flat band della cella è più elevata e di conseguenza anche i campi elettrici sono più alti durante la fase di programmazione [9].

Per quanto riguarda l'interfaccia nitruro-ossido-substrato la cella TANOS si comporta in modo del tutto analogo alla cella SONOS precedentemente descritta: programmazione e cancellazione avvengono per effetto tunnel Fowler-Nordheim.

2.3 Post floating gate

L'esigenza di risolvere le problematiche conseguenti allo scaling dei dispositivi ha condotto la ricerca verso lo sviluppo nuove architetture di memorie con innovativi concetti di funzionamento. In questo contesto va intesa l'idea di sfruttare anche la terza dimensione per integrare un maggior numero di transistors ed incrementare la capacità di memorizzazione dei dati. La possibilità di fabbricare transistors su uno strato di silicio-ossido-silicio, unitamente all'idea di costruire le celle di memoria

"stacked" su più livelli (chiamati *layers*) ha portato allo studio e allo sviluppo di tecnologie per memorie 3D NAND sviluppate verticalmente.

2.3.1 Silicon On Insulator

Negli ultimi anni gli avanzamenti in campo tecnologico nella fabbricazione dei wafers hanno permesso, tra gli altri, lo sviluppo della tecnologia Silicon On Insulator (SOI), la cui prima implementazione industriale fu annunciata nel 1998 [10]. Questi dispositivi sono realizzati come classici transistors il cui substrato non è contattabile direttamente poiché è a contatto con uno strato di ossido (tipicamente diossido di silicio SiO_2) che isola il body della struttura. Una schematizzazione di questo tipo di dispositivo è presentata in figura 2.3.1.

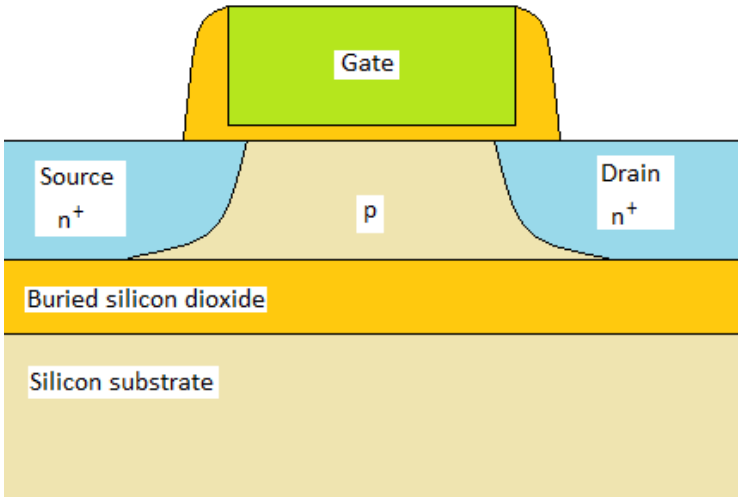


Figura 2.3.1: schema della sezione di un transistor nMOS in tecnologia SOI

Questo tipo di ingegnerizzazione del substrato apporta diversi vantaggi, tra cui la riduzione delle capacità di giunzione (dovuta a una minore area di interfaccia tra l'impianto di source/drain n^+ con il body p), la riduzione del rischio di latch-up nel circuito, l'eliminazione dell'effetto body e una maggiore scalabilità dei circuiti integrati.

La possibilità di realizzare transistors in tecnologia SOI ha dato il via all'ingegnerizzazione della cella fino a raggiungere architetture di celle piuttosto elaborate: è il caso del FinFET. Un transistor così costruito ha la particolarità di avere un sottile strato di silicio (simile a una "pinna") che fa da area attiva in cui si forma il canale di inversione, intorno al quale viene posto il gate ingegnerizzato come quelli discussi nel precedente paragrafo (SONOS o TANOS).

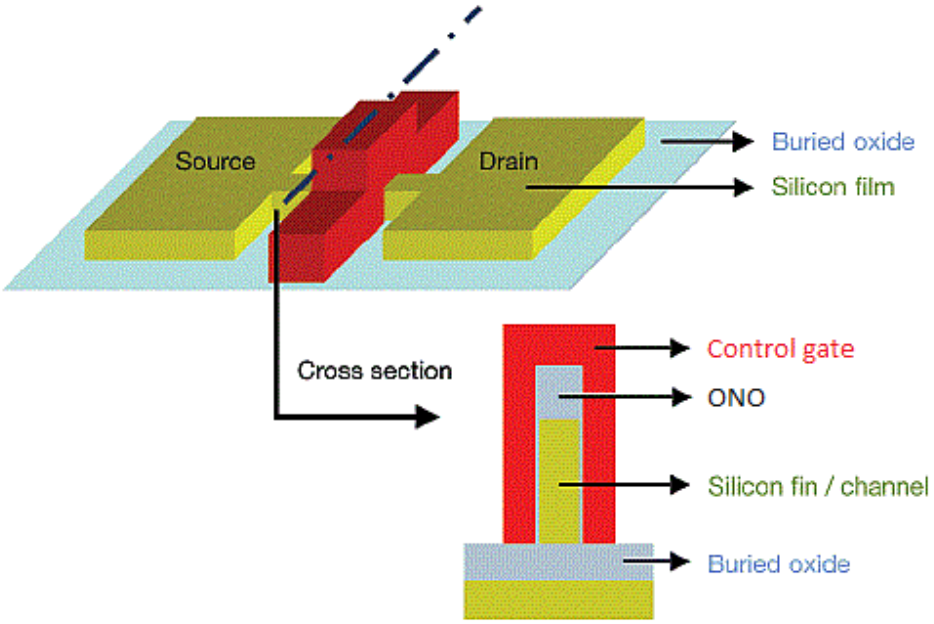


Figura 2.3.2: schema della struttura di un FinFET

Come si vede in figura 2.3.2, la struttura del gate e dello strato ossido-nitruro-ossido avvolge l'area attiva da tre lati: ne risulta un aumento del flusso del campo elettrico tra il substrato e lo strato intrappolante (nitruro) quando viene applicata una polarizzazione al gate o al substrato. L'aumento del campo effettivo si traduce in un aumento della probabilità di tunnel Fowler-Nordheim e incrementa il controllo sul dispositivo cercando di limitare gli effetti di canale corto e le problematiche che sono andate amplificandosi via via che le dimensioni dei dispositivi sono diminuite con lo scaling.

2.3.2 Memorie 3D

Lo sviluppo di memorie verticali si è proposto come valida evoluzione delle classiche NAND planari. La ricerca di strutture e architetture innovative per lo sviluppo di memorie flash NAND ha portato all'idea di creare le stringhe di celle di memoria su diversi piani: in questo caso si parla di memorie *simply stacked NAND* (figura 2.3.3)

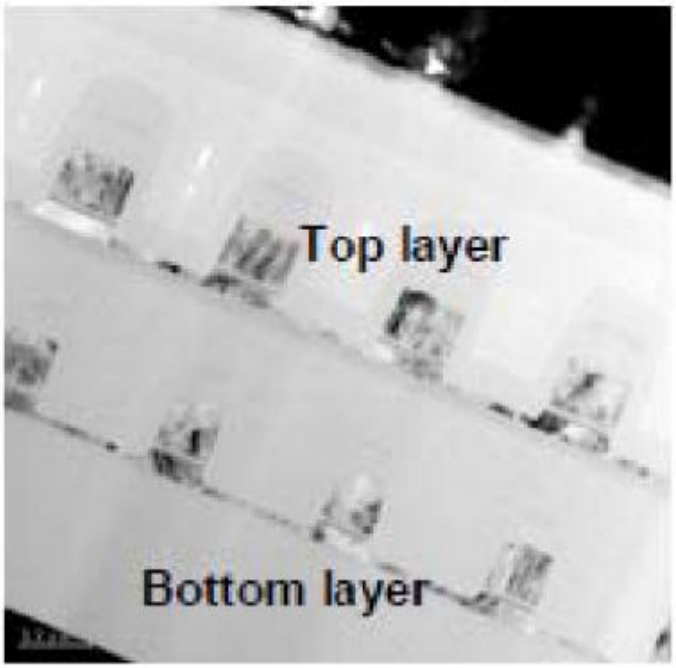


Figura 2.3.3: sezione della struttura di una memoria *simply stacked* osservata al microscopio elettronico. I layers sono realizzati con celle SONOS a body flottante in tecnologia TFT (Thin Film Transistors) [11].

Ulteriori sviluppi di memorie 3D hanno portato a ideare strutture innovative, chiamate *not simply stacked NAND*. Per esempio, il caso delle memorie flash BiCS (Bit Cost Scalable) che realizzano dispositivi di memoria ad altissima densità di immagazzinamento dati a costi ridotti, sfruttando il processo "punch and plug" (figura 2.3.4) .

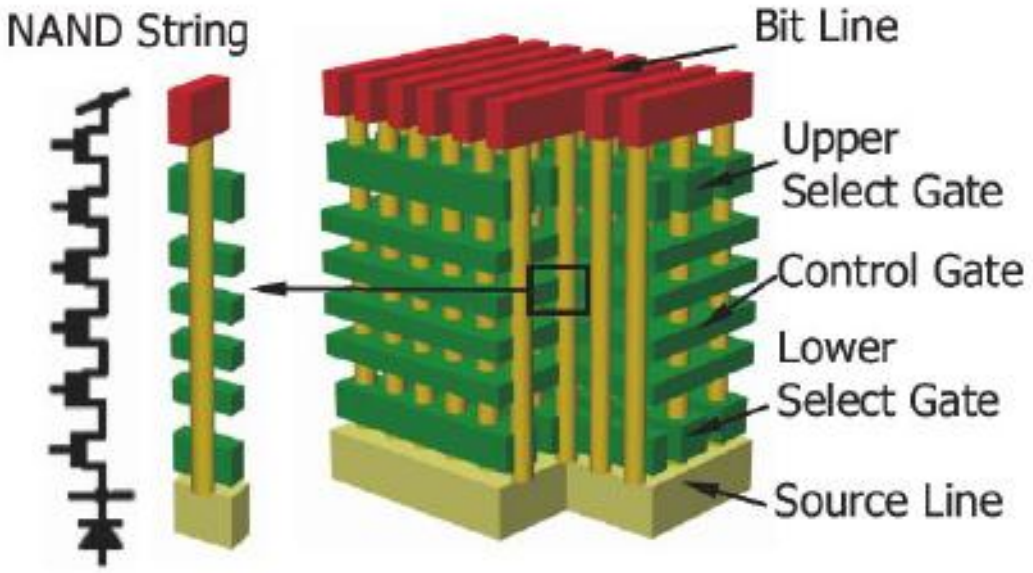


Figura 2.3.4: schema della struttura di una memoria flash BiCS [12].

Un altro esempio di memoria 3D è la VSAT (Vertical Stacked Array Transistors) per la quale le celle di memoria composte da una struttura "double gate" e sistemate una sopra all'altra in verticale, come in figura 2.3.5.

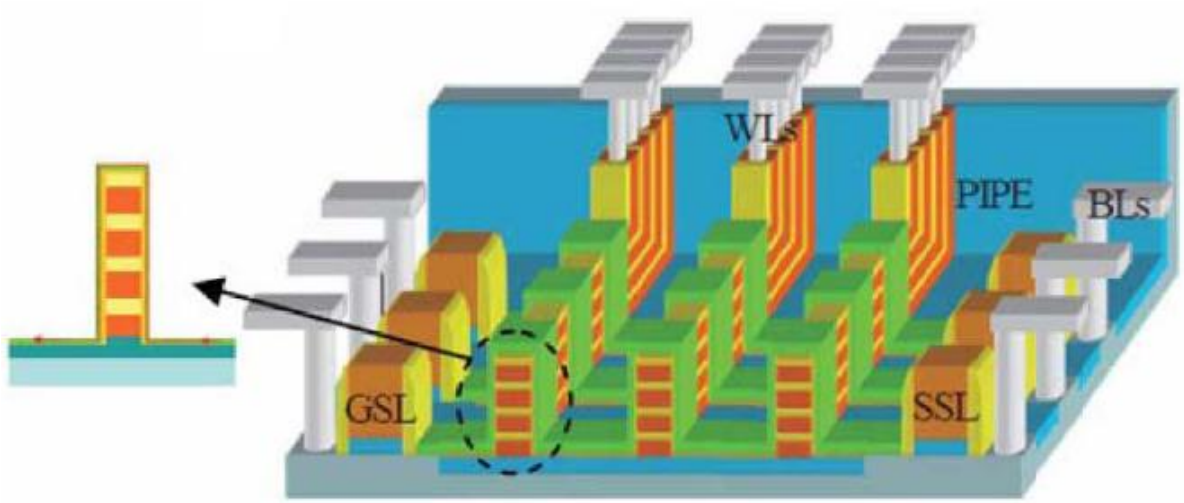


Figura 2.3.5: schema della struttura di una memoria VSAT [13].

Un altro esempio di struttura 3D, schematizzato in figura 2.3.6, è la VG-NAND (Vertical Gate - NAND), nella quale le stringhe sono impilate verticalmente e condividono un gate verticale. Le celle vengono selezionate tramite un'opportuna

polarizzazione delle SSL. Un sorprendente vantaggio di questa tecnologia sta nel fatto che il numero di step di processo necessari per fabbricare una VG-NAND è comparabile con quelli di una classica NAND planare.

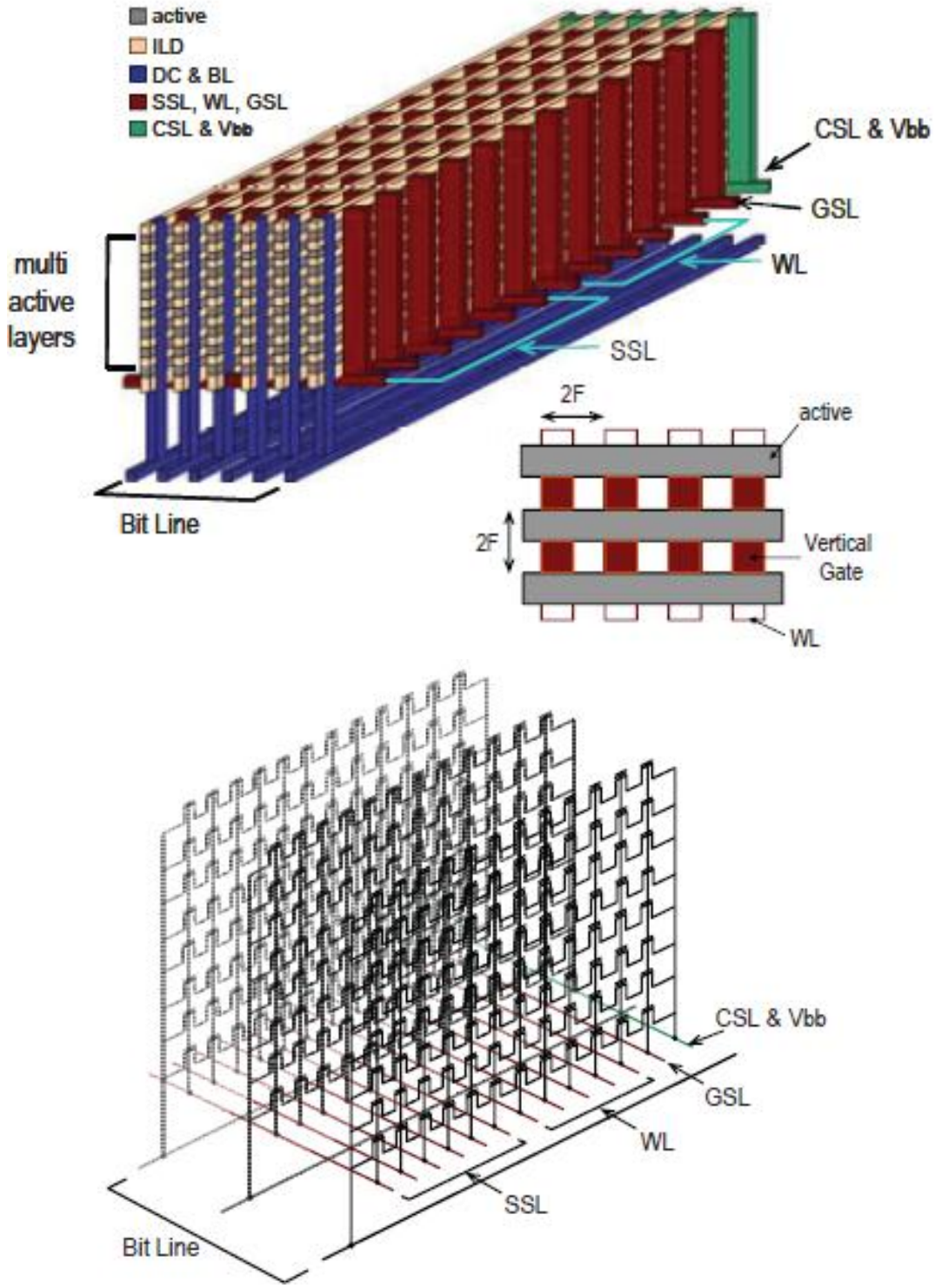


Figura 2.3.6: schemi della struttura di una memoria VG-NAND [14].

2.4 Conclusioni

In questo capitolo abbiamo introdotto il concetto di scaling come linea guida per lo sviluppo tecnologico e abbiamo visto quali problematiche si incontrano procedendo verso una riduzione delle dimensioni della cella di memoria. Per questo la ricerca si è spinta verso la realizzazione di celle di memoria in cui la stack della floating gate è stata ingegnerizzata con differenti materiali e strutture, e al posto delle memorie planari ha preso piede lo studio e la realizzazione di memorie stacked sviluppabili verticalmente. Per realizzare questi array di celle di memoria 3D spesso vengono utilizzati i transistori MOSFET SOI grazie all'elevata scalabilità.

E' quindi fondamentale analizzare le prestazioni e le caratteristiche delle singole celle dallo stack ingegnerizzato e studiare il comportamento delle celle SONOS e TANOS quando il loro body non è contattabile, obiettivo che conseguiremo nei prossimi capitoli.

Capitolo 3

Celle di memoria TANOS standard a body contattabile

3.1 Struttura

I dispositivi di memoria studiati e misurati sperimentalmente in questo lavoro di tesi sono stati progettati e realizzati presso i laboratori Numonyx/Micron ad Agrate Brianza. I campioni misurati sono celle TANOS in tecnologia 52nm realizzate su substrato di silicio monocristallino tipo p con regioni laterali n^+. Il gate di questi dispositivi è costituito da uno stack di TaN (nitruro di tantalio), uno strato di Al_2O_3 (allumina), uno strato di SiN (nitruro di silicio di trapping) e uno strato di SiO_2 (ossido di silicio). La struttura è stata sviluppata su uno strato di silicio policristallino contattabile. Per analizzare l'influenza dello spessore di questi strati sulle prestazioni della cella e della memoria sono stati realizzati diversi wafers con strutture differenti tra loro dal punto di vista morfologico, ovvero modificando gli spessori, e strutturale, ovvero modificando il tipo di nitruro. In figura 3.1.1 sono specificati gli spessori nominali dei vari campioni analizzati per una cella TANOS a body contattabile.

bottom oxide	nitruro	allumina
SiO_2 4,5nm	SiN LPCVD 6nm	Al_2O_3 15nm÷17,5nm

Figura 3.1.1: spessori nominali dei dispositivi TANOS a body contattabile.

Come si può vedere, lo spessore dell'ossido di bottom è stato scelto per garantire accettabili prestazioni di ritenzione della carica intrappolata nel nitruro e per garantire un'adeguata finestra in fase di programmazione e cancellazione, mentre lo spessore dell'allumina è stato scelto all'interno di un range accettabile per diminuire le probabilità di iniezione di carica tra gate e nitruro e viceversa (come si vedrà nel prossimo capitolo, sarà possibile confrontare celle standard a body contattabile e celle SOI a body flottante con spessori di allumina molto simili). Lo strato di nitruro è stato realizzato con LPCVD, si tratta di un nitruro stechiometrico.

La caratterizzazione sperimentale delle celle è stata effettuata tramite probe-station misurando strutture sui wafer di silicio costituite da una stringa di transistors: in questo array le celle vengono selezionate in modo differente a seconda del tipo di operazione da eseguire, indirizzando opportunamente le bit lines (BL) e le word lines (WL). Di seguito, in figura 3.1.2, è riportato uno schema che esemplifica la struttura della stringa (struttura tipica delle memorie NAND). La differenza sostanziale con le classiche architetture NAND risiede nei selettori (selettore di drain DSL e selettore di source SSL) che possiedono nello stack del loro gate il nitruro di trapping e per questo possono essere a loro volta programmati e cancellati pur non essendo delle vere e proprie celle. Naturalmente questa differenza è stata presa in considerazione in tutte le caratterizzazioni mostrate in seguito.

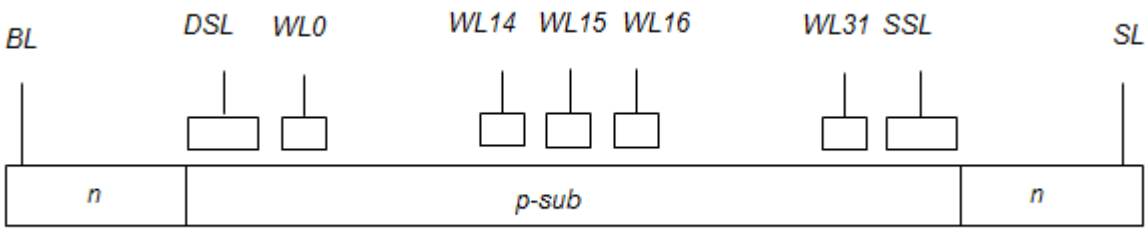


Figura 3.1.2: schema della struttura della stringa NAND

All'applicazione della polarizzazione del gate si misura la corrente e si va a ricostruire la caratteristica I-V della cella, ovvero la *transcaratteristica*. Da questa è possibile estrarre la tensione di soglia misurando il valore di tensione ad una certa corrente (nel caso delle misure effettuate, 10nA). I range di tensione utilizzati per misurare le

caratteristiche di programmazione e cancellazione vanno da 16V a 20V mentre i tempi per i quali viene applicato l'impulso sono cumulativi da 1us a 100ms (1s per la cancellazione).

3.2 Curve di programmazione

Come specificato anche nei capitoli precedenti, la programmazione di una cella TANOS avviene per effetto tunnel di tipo Fowler-Nordheim attraverso l'ossido di tunnel tramite l'applicazione di un campo elettrico elevato ai capi della cella selezionata. Tipicamente si utilizza una tensione di 18V applicata sul contatto della WL da programmare, lasciando a massa BL e SL; le altre WL non selezionate vengono polarizzate ad una tensione V_{pass} di 8V per evitare che vengano programmate e per minimizzare i disturbi sulle
celle adiacenti; i selettori di source e di drain SSL e DSL vengono polarizzati a 3V. Prima di iniziare ogni misura di programmazione a tempo crescente è stata effettuata una cancellazione in modo da svincolare i risultati dal valore iniziale della soglia della cella.

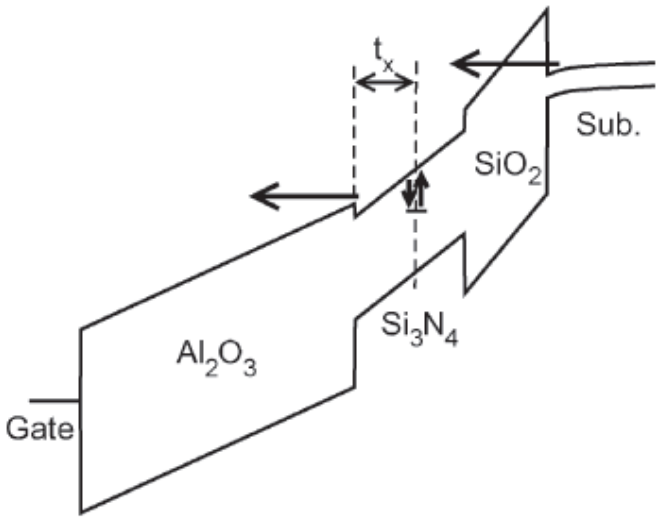


Figura 3.2.1: struttura a bande di una cella TANOS in fase di programmazione; sono schematizzati i flussi di corrente entrante ed uscente dal nitruro e la cattura/rilascio dei portatori ad opera delle trappole [15].

La carica che si accumula nel canale del substrato invertito al di sotto della cella selezionata viene iniettata attraverso l'ossido verso lo strato di nitruro di trapping: assottigliando l'ossido di silicio la probabilità di tunnel aumenta, quindi la cella si programma con maggiore efficienza. Gli elettroni iniettati vengono intrappolati negli stati discreti, andando a modificare l'effettiva tensione di soglia della cella. Esiste tuttavia una probabilità non nulla che gli elettroni presenti nel nitruro vengano iniettati verso il gate metallico: grazie alla presenza dello strato di allumina opportunamente dimensionato si riesce a minimizzare questo effetto mantenendo la carica nello strato di trapping.

E' da notare come in fase di programmazione ed in fase di lettura il substrato non venga polarizzato.

In figura 3.2.2 sono riportate le curve relative alle misure di efficienza di programmazione per celle TANOS a body contattabile con spessore nominale di allumina 15nm; in figura 3.2.3 le curve per celle TANOS con allumina 17,5nm. Le misure sono state effettuate in seguito ad una prima cancellazione in modo da svincolare la cella dal valore di soglia vergine e portarla ad uno stato iniziale cancellato; le curve reiterate a 16V testimoniano la ripetibilità dell'operazione.

E' evidente che per tensioni di comando elevate (20V) e durata degli impulsi lunga (100ms) la differenza tra lo stato iniziale cancellato della cella e lo stato programmato rimane inferiore ai 4V. L'efficienza di programmazione è definita come differenza tra la soglia programmata (risultante dall'operazione) e l'iniziale soglia cancellata.

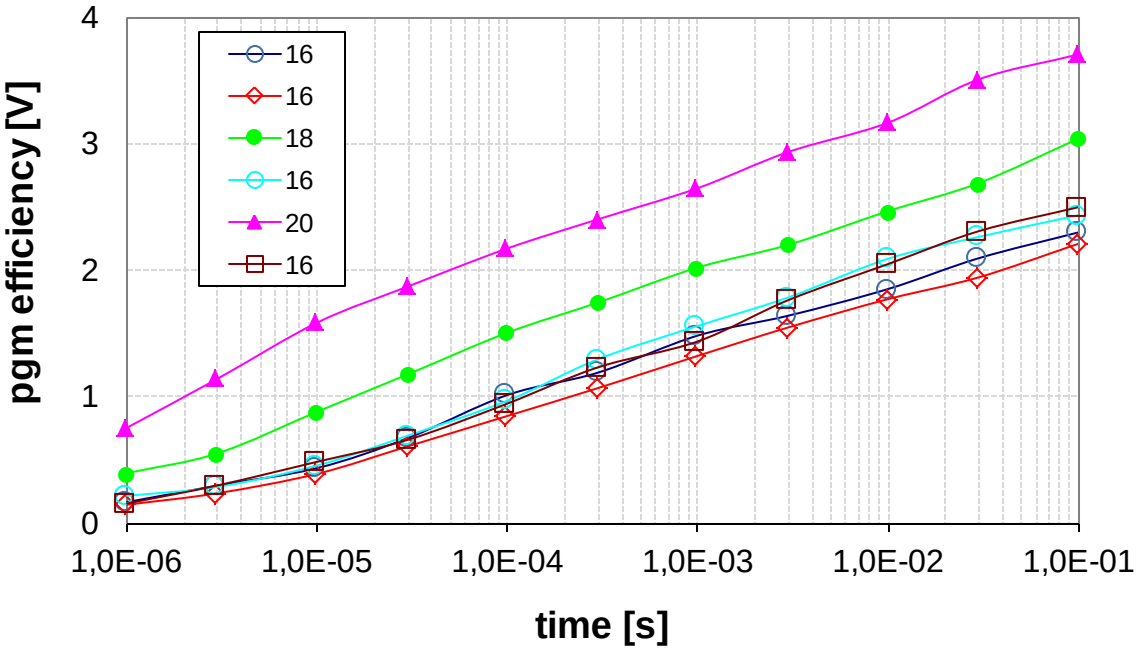


Fig. 3.2.2: curve di efficienza di programmazione di celle TANOS con Al_2O_3 = 15nm.

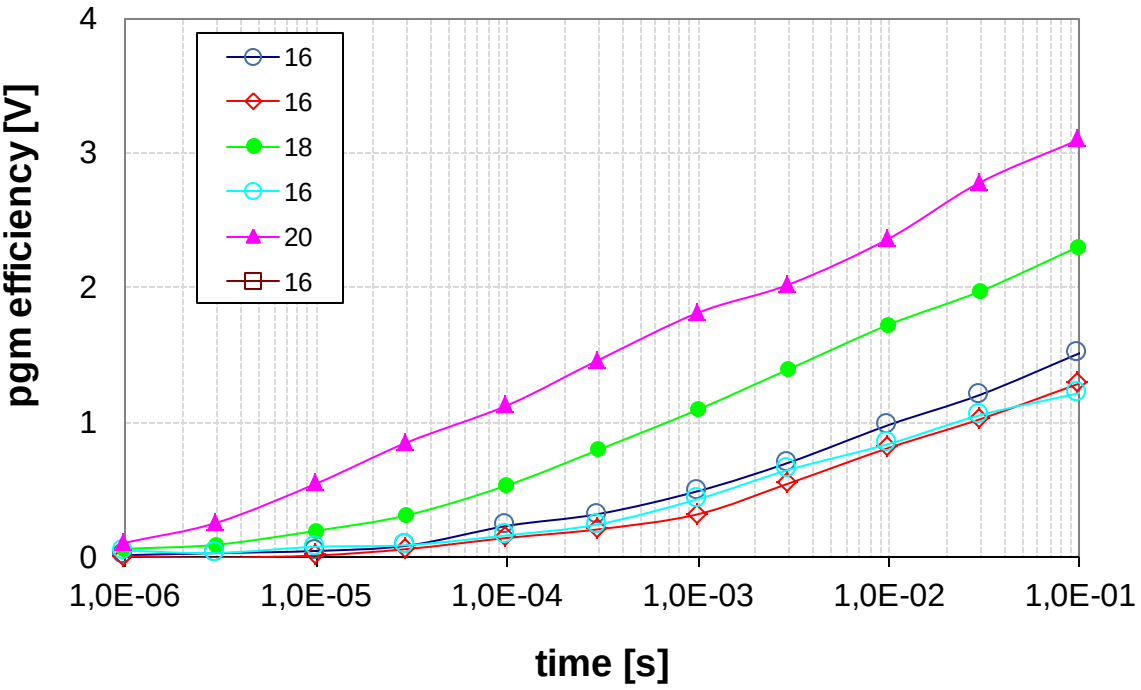


Fig.3.2.3: curve di efficienza di programmazione di celle TANOS con Al_2O_3 = 17,5nm.

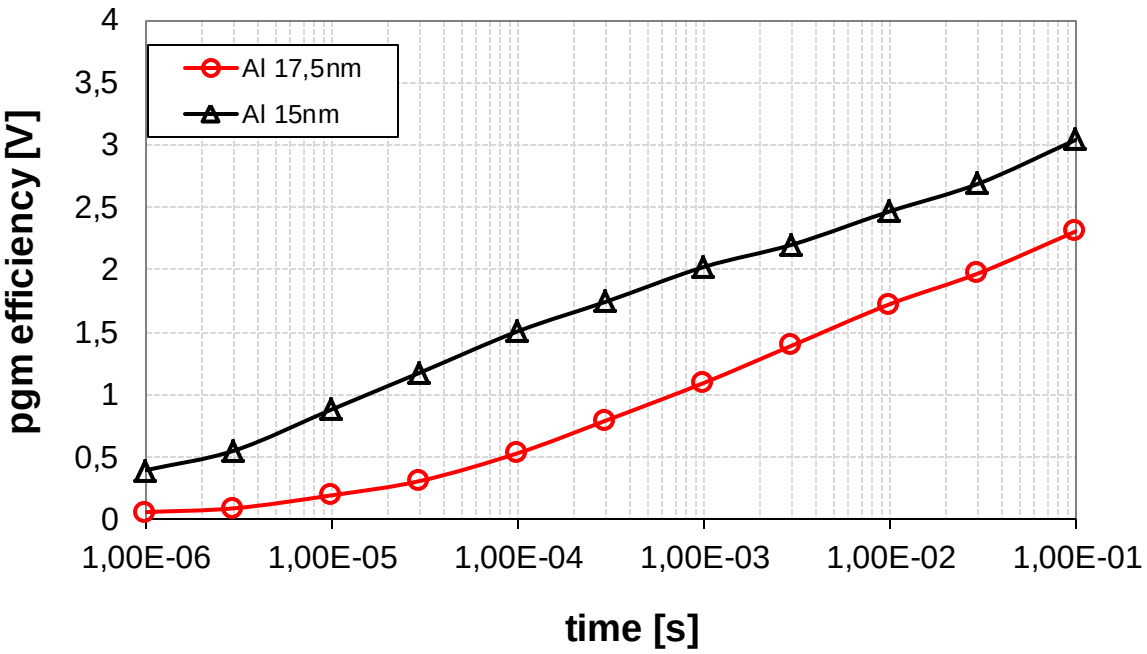


Fig. 3.2.4: confronto tra efficienze di programmazione di celle TANOS con differenti spessori di allumina (tensione di comando 18V).

Il confronto a parità di condizioni di programmazione in figura 3.2.4 mette in luce il fatto che celle con un minor spessore dell'allumina hanno efficienza di programmazione maggiore: questo perché riducendo lo spessore si ha un aumento del campo ai capi dell'ossido di bottom, quindi una maggiore iniezione di carica che viene intrappolata nel nitruro.

3.3 Curve di cancellazione

Come visto nei precedenti capitoli, la cancellazione di una cella di memoria TANOS a body contattabile viene effettuata applicando un'elevata tensione al substrato che crea un campo elettrico tale da favorire l'emissione di elettroni dal nitruro attraverso l'ossido di bottom per effetto tunnel Fowler-Nordheim.

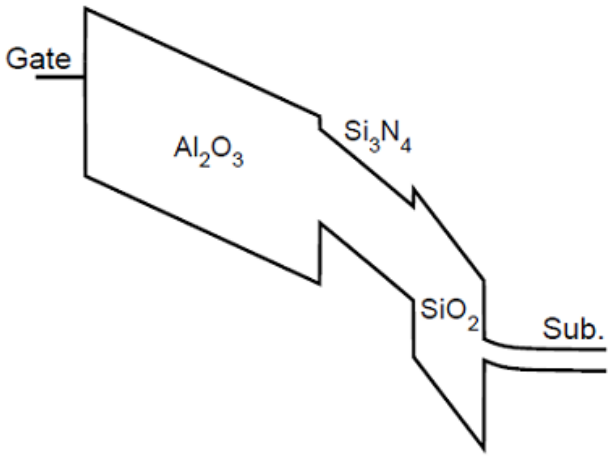


Figura 3.3.1: struttura a bande di una cella TANOS in fase di cancellazione.

L'applicazione del potenziale piega le bande come indicato in figura 3.3.1 e questa situazione favorisce l'emissione delle cariche dallo strato di trapping verso il substrato. L'estrazione della carica dalle trappole dello strato di nitruro comporta una diminuzione della tensione di soglia effettiva della cella: anche in questo caso assottigliando l'ossido viene migliorata l'efficienza con cui gli elettroni vengono emessi. Oltre a questo flusso di elettroni sono presenti altri due contributi, ovvero il flusso di lacune che dal substrato vengono iniettate nel nitruro e il flusso di elettroni che dal nitruro vengono iniettati verso il gate. Il contributo dell'iniezione di lacune diviene via via più importante man mano che l'ossido viene ridotto.

Si è visto nel precedente capitolo come la principale problematica legata alle celle di memoria SONOS sia *l'erase saturation*, che limita i valori di tensione di soglia cancellata della cella a valori anche maggiori rispetto a quelli della cella vergine (o neutra): questo è il motivo che ha condotto la ricerca a sviluppare le celle TANOS. Queste nuove tecnologie attenuano ma non sono immuni alle problematiche relative alla cancellazione. Le origini dell'*erase saturation* sono da ricercarsi nel contributo accennato poco sopra dovuto all'iniezione di elettroni per effetto tunnel dal gate al nitruro. La presenza dell'allumina funge da schermo tanto più efficace quanto più l'allumina è spessa, tuttavia una parte di questo contributo è osservabile anche per le memorie TANOS. Quando il flusso di cariche iniettato dal gate bilancia la diminuzione di carica nel nitruro attraverso l'ossido di bottom (sia per emissione di elettroni che per iniezione di lacune) la tensione di soglia della cella non subisce più

sostanziali modifiche, l'effetto di cancellazione sulla cella non è più efficace e la tensione di soglia raggiunge un valore di saturazione. Affinché i tre contributi di corrente si bilancino è necessario che i campi elettrici diano luogo ad un flusso di carica netta pari a zero: questo spiega il fatto che il valore della tensione di soglia a cui si verifica questo fenomeno dipende dalla tensione applicata alla struttura poiché aumentare la caduta di tensione significa aumentare i campi elettrici e quindi raggiungere prima il bilancio delle correnti e anche il livello di soglia a cui si verifica la saturazione.

In figura 3.3.2 e in figura 3.3.3 sono mostrate le curve delle misure di efficienza di cancellazione per celle TANOS a substrato contattabile con spessore nominale di allumina rispettivamente di 15nm e 17,5nm. Prima di effettuare queste misure è stato applicato un impulso di programmazione alla cella in modo da aumentare la sua tensione di soglia e valutare l'efficienza come differenza tra la soglia cancellata e quella dello stato iniziale programmato. Per elevate tensioni di comando (20V) ed ampie durate degli impulsi (100ms÷1s) si può notare l'effetto di appiattimento delle curve dovuto all'*erase saturation*; le efficienze di queste curve non superano i -6V.

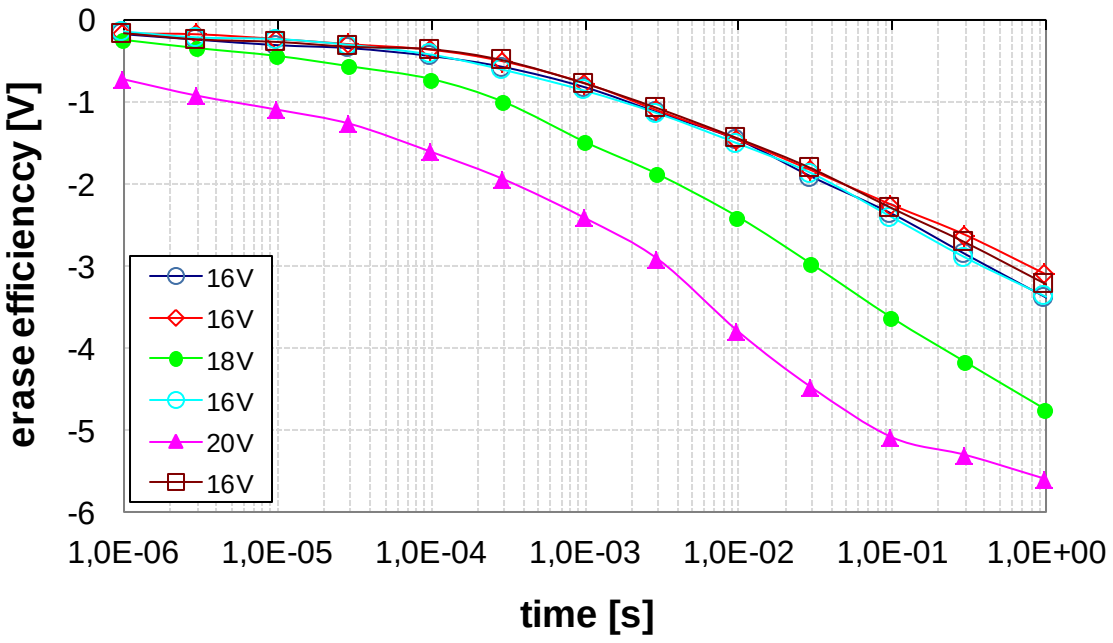


Figura 3.3.2: curve di cancellazione di celle TANOS con Al_2O_3 = 15nm.

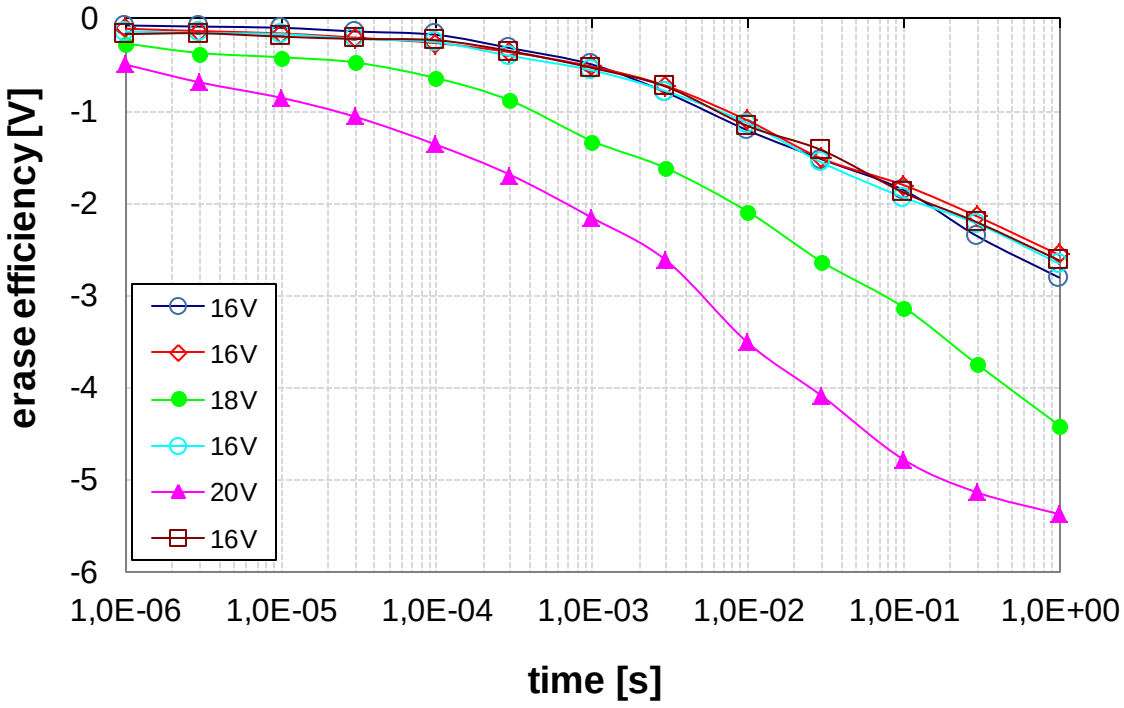


Figura 3.3.3: curve di cancellazione di celle TANOS con Al_2O_3 = 17,5nm.

Il confronto presente in figura 3.3.4 mostra che, in modo simile a quanto si verifica nell'operazione di programmazione, assottigliare l'allumina significa diminuire lo spessore equivalente dell'ossido dello stack e corrispondentemente aumentare il campo elettrico ai capi dell'ossido di bottom. Per questo a una diminuzione dello spessore dell'allumina da corrisponde un aumento dell'efficienza di cancellazione, oltre che di programmazione (vedi paragrafo precedente).

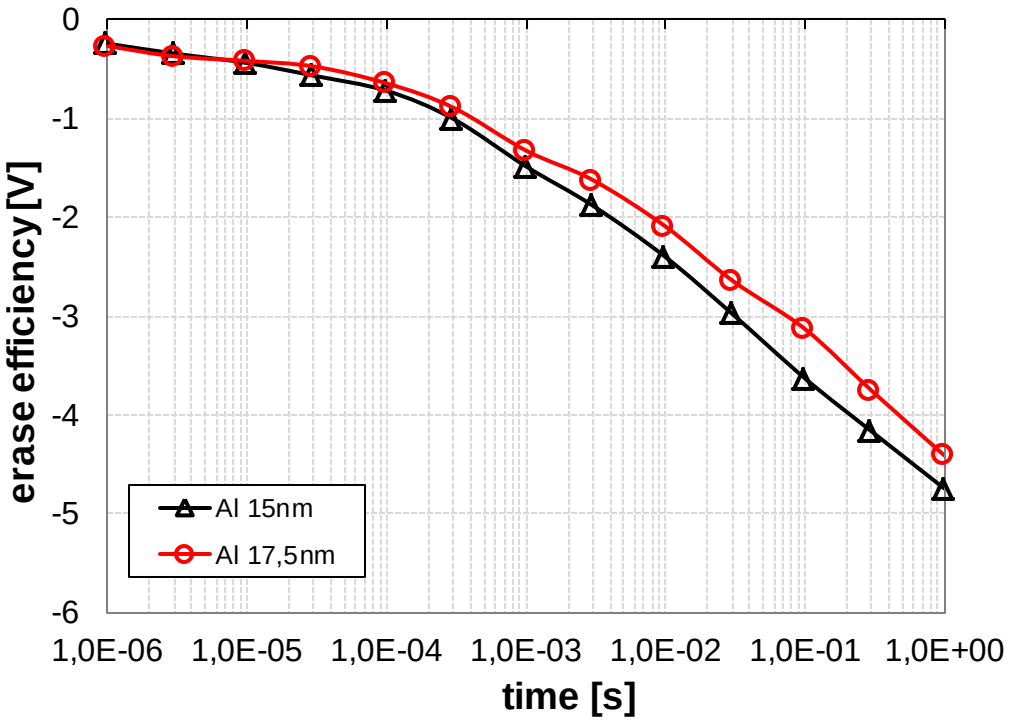


Fig. 3.3.4: confronto tra efficienze di cancellazione di celle TANOS con differenti spessori di allumina (tensione di comando 18V).

3.4 Ciclatura

Durante ogni operazione di programmazione e cancellazione di una memoria TANOS si ha un flusso di cariche (elettroni e lacune) che attraversa l'ossido di bottom. Il passaggio di questa corrente di tunnel è un effetto distruttivo per il quale gli urti delle cariche con il reticolo creano difetti nell'ossido. Queste imperfezioni sono responsabili del degrado della cella poiché la regolarità e la purezza del reticolo dell'ossido di bottom sono parametri critici che influenzano le operazioni stesse del transistor.

Contrariamente ad una classica cella a floating gate, in cui le operazioni di programmazione e cancellazione coinvolgono flussi iniettati ed emessi di soli elettroni, in una cella TANOS in fase di cancellazione è presente anche un flusso di lacune. Questo effetto porta ad una differente influenza delle imperfezioni dell'ossido nell'evoluzione della tensione di soglia in ciclatura in dipendenza dall'operazione che viene compiuta, programmazione o cancellazione [16]. Mentre una cella floating gate

mostra un andamento asimmetrico con un degrado maggiore in cancellazione, una TANOS presenta uno shift simmetrico di due livelli di soglia, come mostrato in figura 3.4.1.

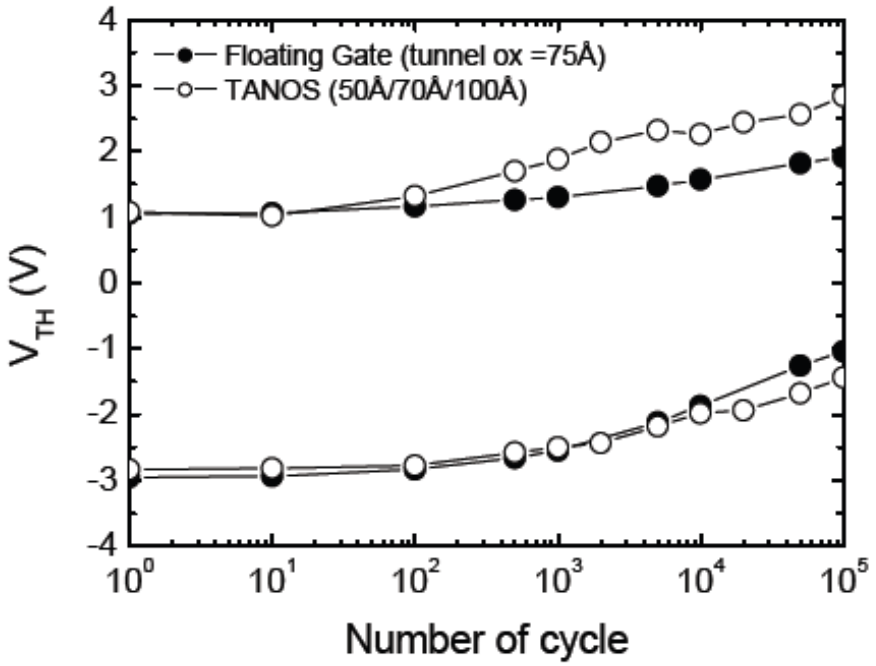


Figura 3.4.1: confronto tra finestre di ciclatura di una floating gate e di una TANOS [15]

Per valutare l'effetto dell'aumento del numero di imperfezioni dell'ossido e del degrado in ciclatura si prendano in considerazione l'evoluzione delle curve transcaratteristiche I-V parametrizzate in funzione del numero di cicli cui è stata sottoposta la cella in figura 3.4.2.

Come si può notare, le curve sottoposte ad un certo numero di cicli (nel caso in figura, circa 3000) sono all'incirca repliche traslate della curva al primo ciclo, il che è indice di riproducibilità delle operazioni di programmazione e cancellazione senza alterare eccessivamente le prestazioni della cella; per numero di cicli via via maggiore si ha un piegamento delle curve con corrispondente peggioramento della pendenza di sottosoglia e diminuzione di guadagno differenziale e corrente di saturazione.

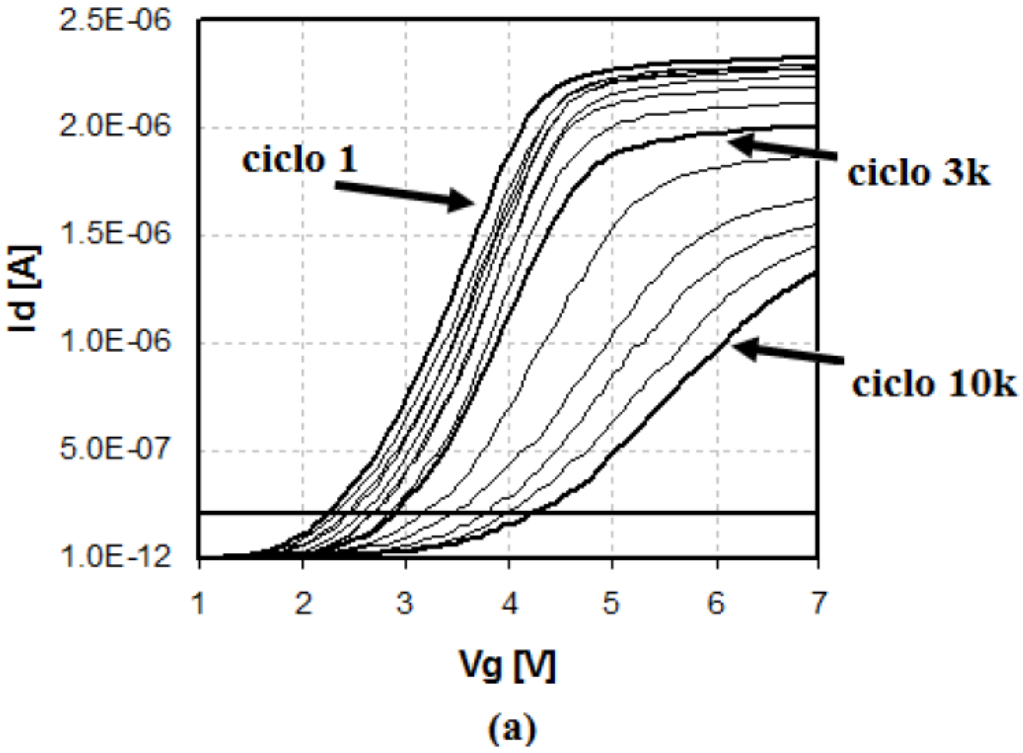


Figura 3.4.2: transcaratteristiche I-V dello stato programmato di una cella TANOS in funzione del numero di cicli; in scala lineare (a) si notano la diminuzione del guadagno differenziale e della corrente di saturazione; in scala logaritmica (b) si nota l'aumento di pendenza del sottosoglia [16].

3.5 Ritenzione

Nel secondo capitolo è stato presentato il concetto di ritenzione, ovvero la capacità di una memoria di mantenere inalterato il dato memorizzato per un certo periodo di tempo lasciando tutti i suoi terminali flottanti (in condizioni di storage), caratteristica critica per valutare l'affidabilità di una memoria non volatile. Durante la ritenzione non vi è tensione applicata e i terminali della struttura vengono lasciati flottanti, tuttavia è presente una differenza di potenziale dovuta alla carica intrappolata presente nel nitruro. I campi elettrici che si sviluppano sono comunque molto piccoli rispetto a quelli presenti in fase di programmazione e cancellazione, quindi la carica persa attraverso l'ossido di bottom segue principalmente le regole del tunnel diretto che dipendono soprattutto dallo spessore della barriera tra ossido e substrato; inoltre, ridurre lo spessore di bottom significa aumentare i campi elettrici ai capi dell'ossido. Le evidenze sperimentali hanno confermato il fatto che riducendo lo spessore di ossido la perdita di carica in ritenzione aumenta. Questo andamento si scontra con quello visto per l'efficienza di programmazione e (soprattutto) cancellazione discusso in precedenza, secondo cui un assottigliamento dell'ossido di bottom porta miglioramenti all'efficienza di programmazione e di cancellazione: il trade-off tra le prestazioni in ritenzione e l'efficacia con cui una cella viene cancellata e programmata dipende direttamente dallo spessore dell'ossido di bottom. La criticità di questo parametro evidenzia come sia necessario realizzare un ossido di bottom con il minor numero di difetti possibile, poiché le cariche intrappolate possono sfruttare gli stati messi a disposizione dalle imperfezioni per superare la barriera dell'ossido e incrementare la perdita di carica nel tempo.

Le celle vergini sulle quali sono state condotte le misure sperimentali di ritenzione sono state inizialmente programmate con impulsi da 18V in modo da raggiungere un livello operativo che corrisponde ad un salto di soglia all'incirca di 3,5V rispetto alla soglia vergine. I wafer sono successivamente stati inseriti in un forno a temperatura 60°C (bake) per agevolare l'emissione delle cariche intrappolate nel nitruro per emissione termoionica. A intervalli di tempo in scala logaritmica fino a 500 ore sono stati letti i valori delle tensioni di soglia ed estrapolati i salti i soglia delle celle. In base

a questi dati è possibile estrapolare la perdita di carica su un periodo di interesse che solitamente è di 10 anni.

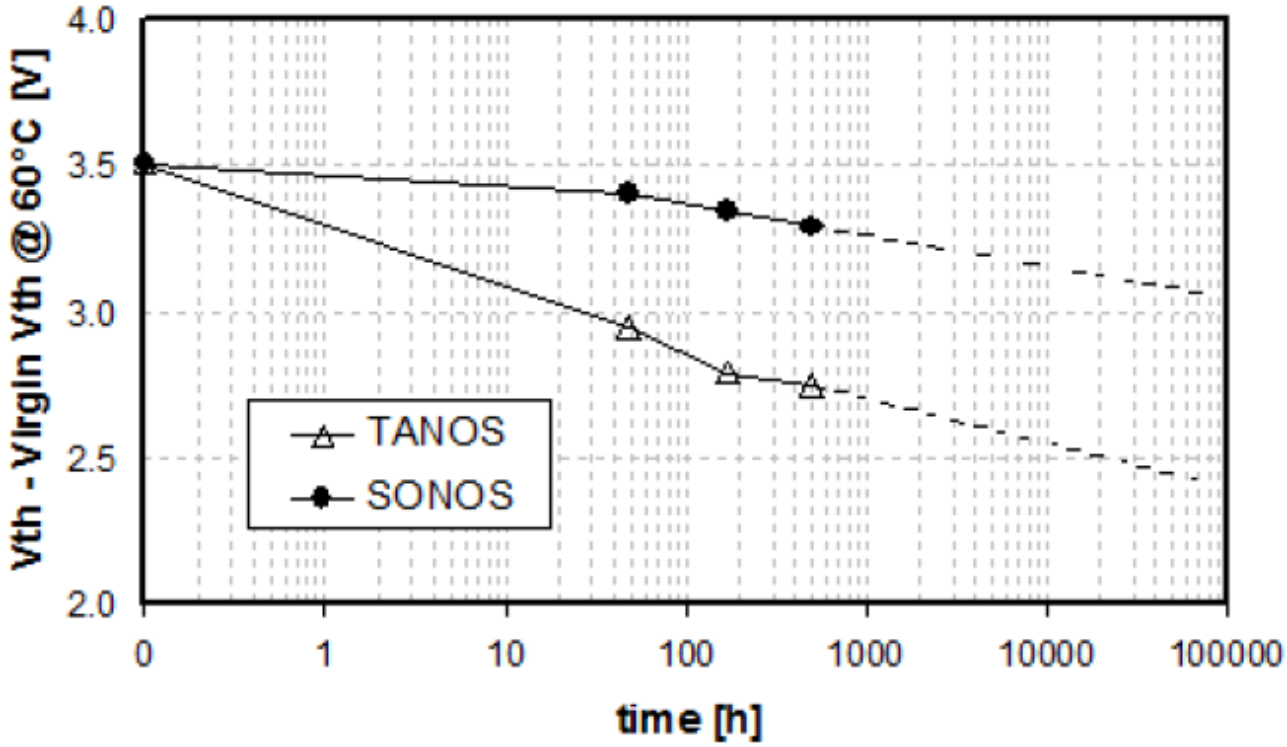


Figura 3.5.1: confronto tra l'evoluzione temporale dello stato programmato normalizzato a 3.5V a 60°C per una cella TANOS e una SONOS con medesimi spessori di bottom e nitruro; come si può vedere la perdita è decisamente inferiore nel caso di una struttura SONOS.

Dalla figura 3.5.1 si può stimare il valore della perdita in ritenzione per celle TANOS standard planari, che alle 60 ore vale più di 0,5V (per le SONOS planari standard vale circa 0,1V). Questi dati verranno utilizzati per un confronto con le architetture che saranno presentate nei successivi capitoli.

3.6 Conclusioni

In questo capitolo abbiamo descritto le strutture TANOS a body contattabile e i metodi con cui sono state effettuate le misure sperimentali, presentando le caratteristiche determinanti per la valutazione delle prestazioni della cella. Abbiamo in particolare introdotto i dati relativi alle celle TANOS con diversi spessori di allumina (Al_2O_3 =

15nm e Al_2O_3 = 17,5nm), esaminando graficamente le caratteristiche di queste strutture. Abbiamo successivamente presentato le prestazioni delle celle TANOS in condizioni di ciclatura e ritenzione, descrivendo i metodi utilizzati per ottenere dati valutativi di queste caratteristiche. Le celle finora descritte e caratterizzate sono celle standard in cui il gate è stato opportunamente ingegnerizzato seguendo l'idea del charge trapping. Nel prossimo capitolo verranno presentate le qualità di queste stesse celle TANOS realizzate però su uno strato di ossido in modo da realizzare transistors SOI: si parla quindi di TANOS floating body.

Capitolo 4

Celle di memoria TANOS Floating Body

4.1 Struttura

Le strutture SOI sono state introdotte nei capitoli precedenti e ne sono state evidenziate le caratteristiche che ne permettono lo scaling. La possibilità dunque di unire i benefici dell'ingegnerizzazione degli strati del gate con la possibilità di creare strutture a body flottante ha portato alla realizzazione di dispositivi TANOS floating body. L'architettura di questi dispositivi è del tutto analoga a quella vista in precedenza: il gate è stato realizzato con uno stack di TaN (nitruro di tantalio), Al_2O_3 (allumina) di spessore 16nm, SiN (nitruro di silicio di trapping) di spessore 6nm e uno strato di SiO_2 (ossido di silicio) di spessore 4,5nm. Questa struttura è realizzata su un substrato di silicio policristallino di spessore nominale di 20nm che a sua volta poggia su uno spesso strato di ossido di silicio: non è dunque possibile contattare direttamente il substrato. Un'affinità con le strutture TANOS a substrato contattabile sta nel fatto che anche in questo caso i selettori DSL e SSL hanno uno stack del gate in cui è presente il nitruro di trapping come per le celle della stringa: a differenza delle classiche architetture NAND, i selettori possono essere programmati e cancellati, caratteristica presa in considerazione nelle successive analisi.

Le misure sono state effettuate tramite probe station contattando i terminali delle stringhe di transistors realizzati su wafer di silicio. Ogni cella è identificata da un contatto di Word Line da 0 a 31, in fase di lettura viene misurata la transcaratteristica della cella interessata e viene estratta la tensione di soglia corrispondente ad un fissato valore di corrente (10nA). I range di tensione di comando per le operazioni di programmazione e cancellazione vanno da 16 a 20 V, mentre i tempi per i quali viene

applicato l'impulso sono cumulativi in scala logaritmica da 1us a 100ms (1s per la cancellazione).

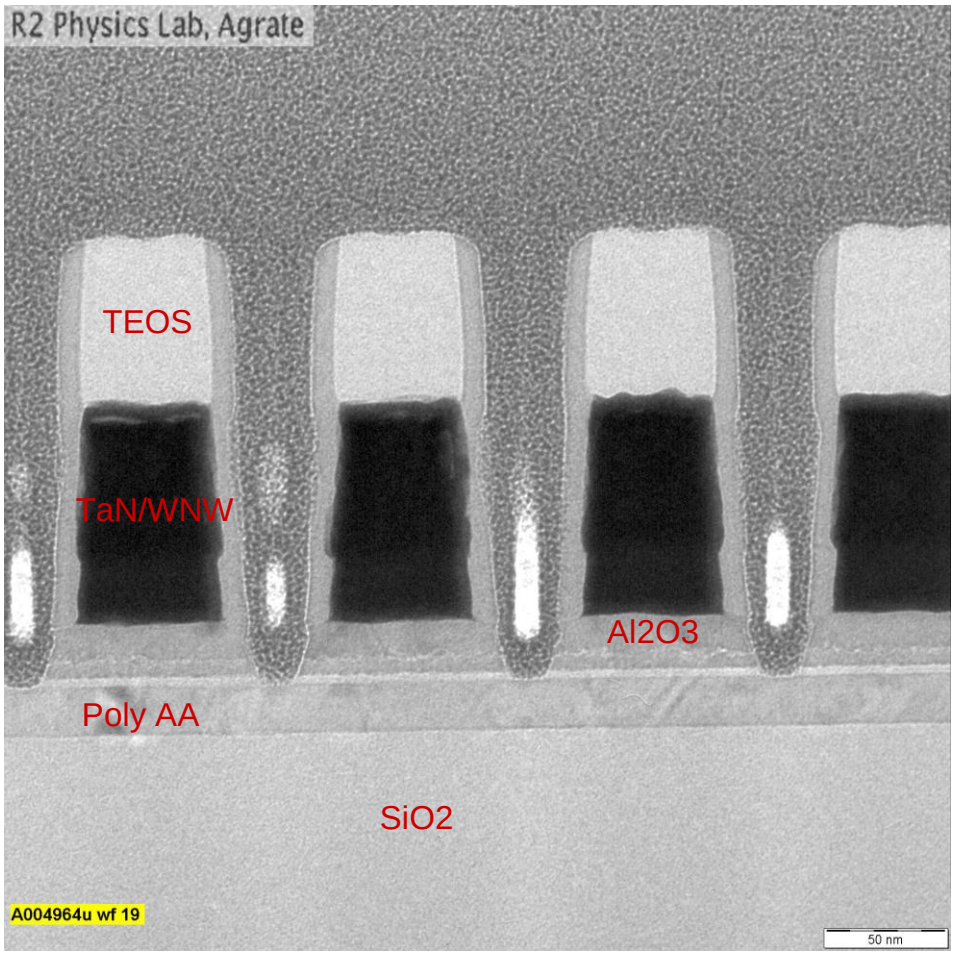


Figura 4.1.1: immagine TEM della struttura della stringa TANOS floating body

Al fine di caratterizzare al meglio le qualità della tecnologia e per cercare soluzioni con prestazioni più spinte, le strutture delle celle e delle stringhe sono state modificate. Sono state condotte misure su wafer con substrato in silicio monocristallino, oltre che policristallino; il nitruro di gate è stato realizzato stechiometrico oppure arricchito, ovvero un nitruro con una percentuale maggiore di silicio.

Nella tabella sottostante sono presentate le caratteristiche tecnologiche dei wafers caratterizzati in questo capitolo.

substrato	bottom oxide	nitruro	allumina
Si p-type 25nm	SiO_2 4,5nm	SiN LPCVD 6nm	Al_2O_3 16nm

Figura 4.1: spessori nominali dei dispositivi TANOS a body flottante.

4.1.1 Stringhe junction e stringhe junctionless

Come descritto nei capitoli precedenti, uno dei vantaggi delle stringhe NAND è la possibilità per i transistors di condividere le zone di source e drain con le celle adiacenti. Un'importante modifica all'architettura della stringa è stata effettuata realizzando array cosiddetti junctionless, che non presentano i classici impianti n^+ nel substrato p sottostante tra una cella e l'altra pur mantenendo inalterato il drogaggio delle zone contattate ai lati della stringa. In questo modo, i transistors non condividono più le zone di source e drain perché queste giunzioni laterali non sono mai state formate. Le stringhe junctionless sono state realizzate in due modi: tramite una maschera che schermi gli impianti di drogante per le singole celle; oppure tramite l'effetto ombra, per mezzo del quale, utilizzando un certo angolo (tilt) per impiantare le cariche che formano le giunzioni tra un transistor e l'altro, i gate dei transistors stessi riescono a oscurare le zone di substrato di interesse.

In fase di lettura, quando viene applicata una polarizzazione alla WL corrispondente alla cella selezionata, oltre alla classica formazione del canale si osserva il fenomeno dell'inversione di carica anche nel substrato tra una cella e l'altra: questo effetto è dovuto agli accoppiamenti capacitivi laterali tra il gate e il substrato e garantisce la conduzione nella stringa in fase di lettura della corrente.

4.2 Curve di programmazione

Dal punto di vista della programmazione, la stringa è stata polarizzata nel modo seguente: è stata applicata la tensione di comando alla WL selezionata mentre tutte le altre WL sono state portate alla tensione V_{pass} di 8V per minimizzare i disturbi sulle celle adiacenti (*boosting*), sufficientemente elevata da permettere la formazione del

canale ma non troppo alta onde evitare programmazioni indesiderate; i selettori di drain (DSL) e di source (SSL) sono stati portati a 3V, la BL è stata posta a 0V mentre la SL a 3V. Si tratta sostanzialmente di una classica programmazione di una cella NAND e da questo punto di vista il body flottante non ha alcun impatto (rimane fisso in tensione a massa). Una volta effettuata la programmazione della cella, si è andati a leggere la transcaratteristica e ad estrapolare la tensione di soglia della cella alla corrente di 10nA.

.

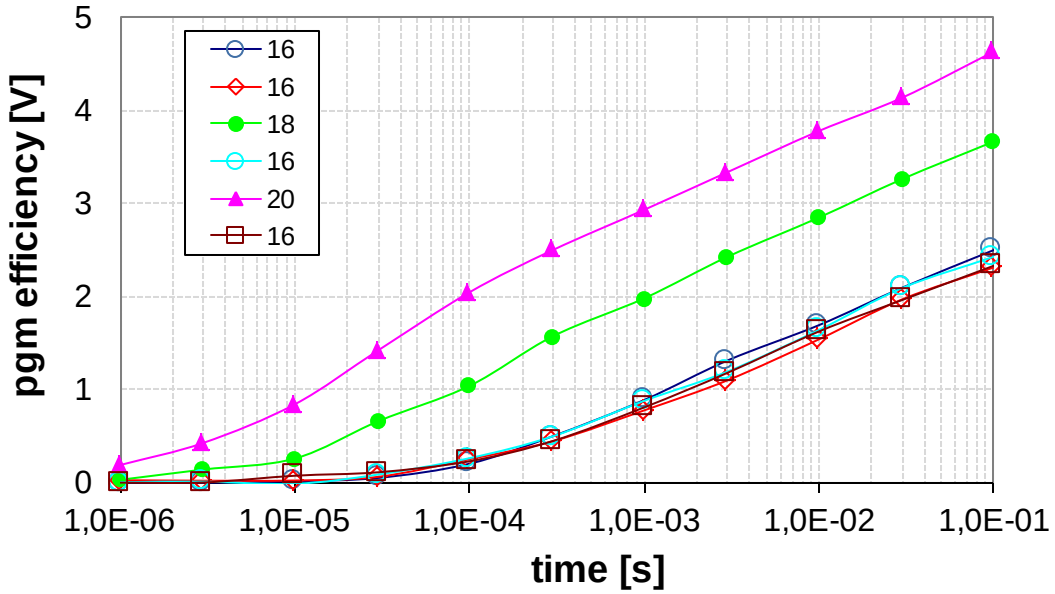


Figura 4.2.1: curve di efficienza di programmazione di celle TANOS floating body con substrato in silicio policristallino e impianti di giunzione.

In figura 4.2.1 sono evidenziate le curve di efficienza di programmazione per una cella TANOS a substrato flottante in funzione della durata dell'impulso e parametrizzate rispetto alla tensione di comando. L'efficienza di programmazione è definita sempre come differenza tra il valore di soglia programmato e quello iniziale cancellato.

La programmazione avviene in modo del tutto analogo alle classiche strutture TANOS planari: in figura 4.2.2 viene presentato un confronto tra le curve di programmazione della struttura floating body appena presentata e delle due strutture planari misurate nel capitolo precedente. Come si può notare, gli andamenti delle curve sono sostanzialmente allineati: questo risultato non è una sorpresa poiché in nessuno dei due casi (planare a floating body) la programmazione comporta l'applicazione di una

tensione di polarizzazione al substrato, pertanto i meccanismi fisici che regolano questa operazione sono del tutto analoghi. In fase di programmazione, la carica si accumula nel canale di inversione del substrato, viene iniettata per effetto tunnel attraverso l'ossido e si intrappola negli stati discreti del nitruro. Si nota infine che i valori di efficienza per elevate tensioni (20V) e lunghi tempi di applicazione dell'impulso (100ms) superano i 4V (a differenza di quanto accade per le TANOS standard) mostrando efficienze anche più elevate del caso planare.

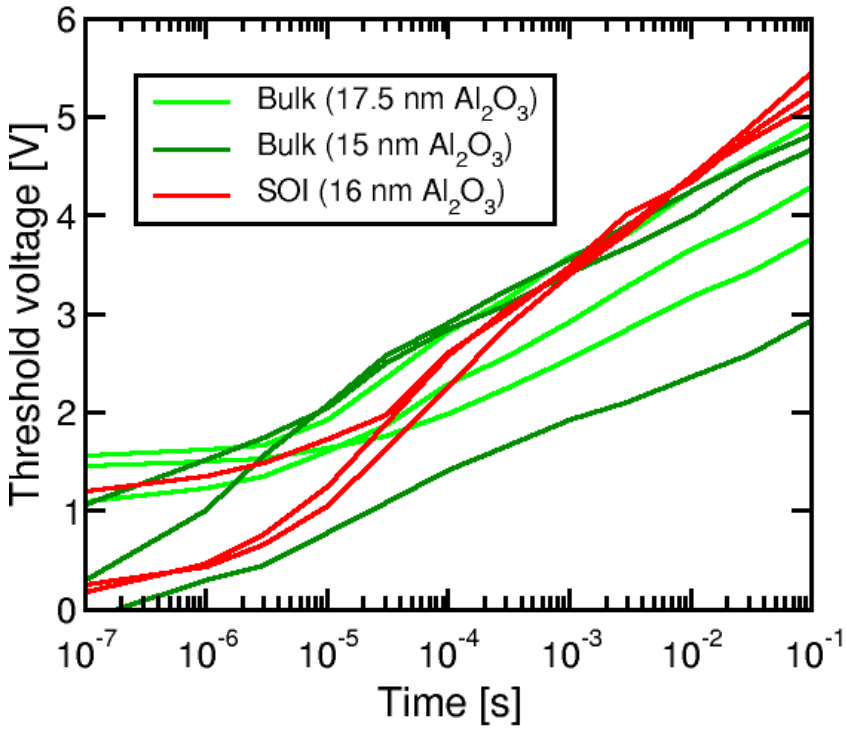
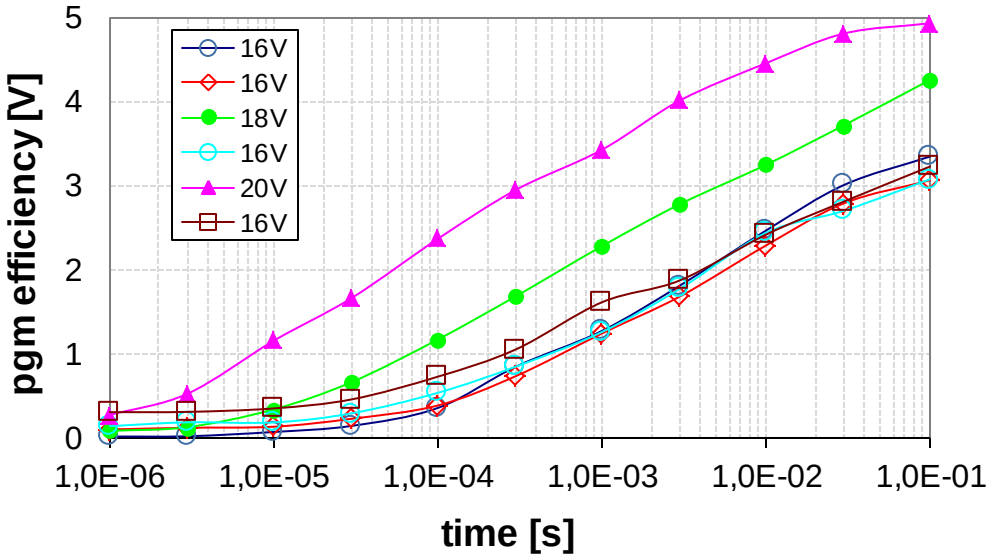


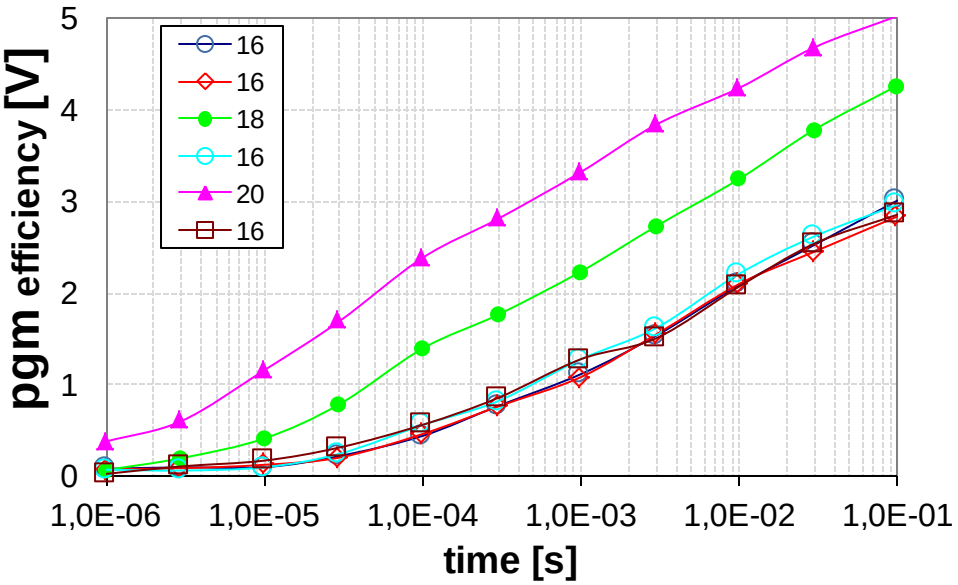
Figura 4.2.2: confronto tra curve di programmazione tra celle TANOS standard e TANOS floating body (tensione di comando 18V).

Un interessante confronto di curve di programmazione riguarda le strutture con e senza le giunzioni laterali precedentemente descritte (junction-junctionless). Abbiamo visto in figura 4.2.1 le curve di efficienza di programmazione per le strutture junction; le curve relative alle strutture junctionless sono presentate in figura 4.2.3. Come si può notare, anche dalla figura 4.2.4 di confronto tra le strutture junction-junctionless, gli andamenti delle curve e i valori di efficienza raggiunti per i diversi istanti di tempo sono del tutto analoghi tra loro; inoltre, gli andamenti delle curve sono simili a quelle delle strutture con impianti di giunzione presenti, mostrando valori di efficienza di

poco più elevati nel caso senza giunzioni. La misura sperimentale ha confermato la possibilità per le stringhe junctionless di avere prestazioni in programmazione simili alle architetture junction: i vantaggi di avere una stringa senza giunzioni si trovano nella semplificazione della struttura della stringa e nella possibilità di scalare ulteriormente le dimensioni dei transistors .



(a)



(b)

Figura 4.2.3.: curve di efficienza di programmazione di celle TANOS floating body con substrato in silicio policristallino e impianti di giunzione mascherati (a) e con impianti tiltati (b).

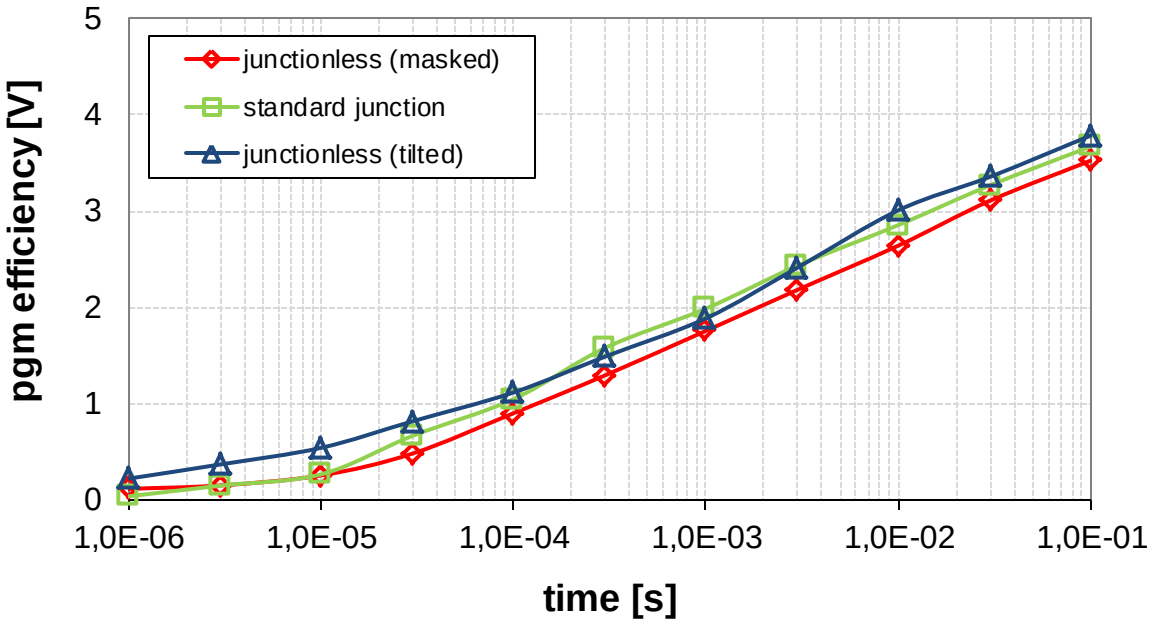


Figura 4.2.4: confronto tra curve di efficienza di programmazione di celle TANOS floating body con e senza impianti di giunzione

Ciò che viene osservato in queste strutture con substrato in silicio policristallino è una rilevante diminuzione della corrente che scorre nei transistors. Per cercare di incrementare la corrente sono state realizzate strutture con substrato in silicio monocristallino. Questa configurazione cristallografica più regolare non presenta i grani caratteristici del policristallo, che possono essere considerati a tutti gli effetti come difetti del reticolo cristallino che ostacolano la propagazione della carica. Avendo una minore difettosità cristallografica, l'utilizzo del monocristallo come substrato ha lo scopo di aumentare la conduttività (e quindi la corrente a parità di tensione applicata) del silicio su cui è costruita la stringa, ovvero i valori di corrente di saturazione rilevati a parità di tensione di polarizzazione del gate.

Dalla figura 4.2.6 si può osservare come i valori di efficienza a parità di durata dell'impulso applicato siano maggiori, seppur di poco, rispetto al caso del policristallo (si vede che lo scarto tra il valore della soglia iniziale e quello finale è leggermente maggiore nel caso del monocristallo).

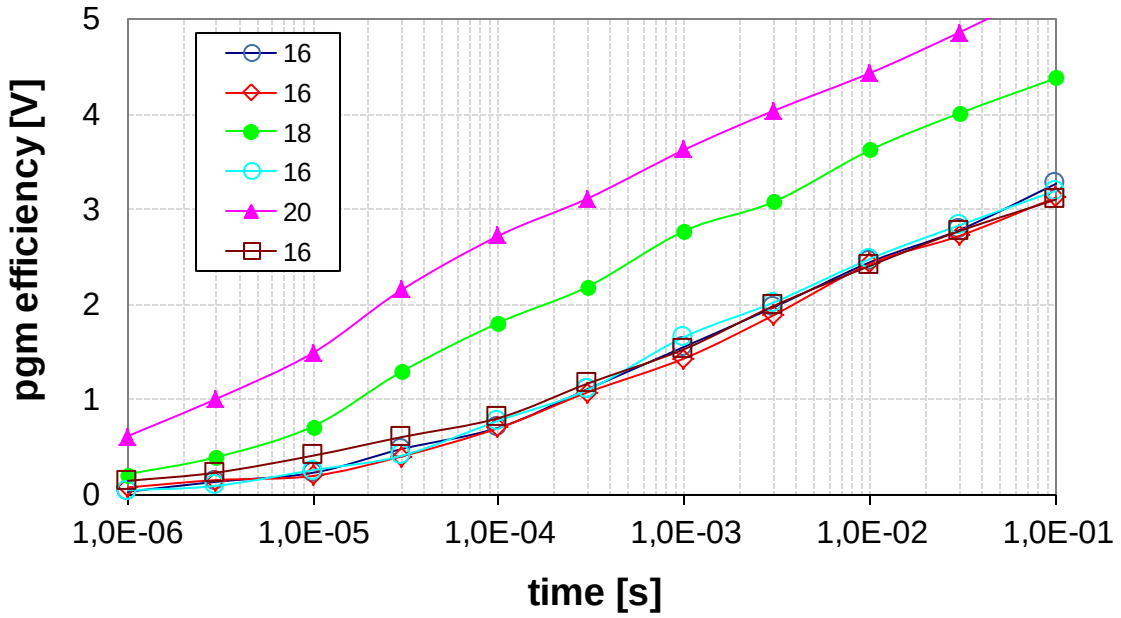


Figura 4.2.5: curve di efficienza di programmazione di celle TANOS floating body con substrato in silicio monocristallino e impianti di giunzione tiltati.

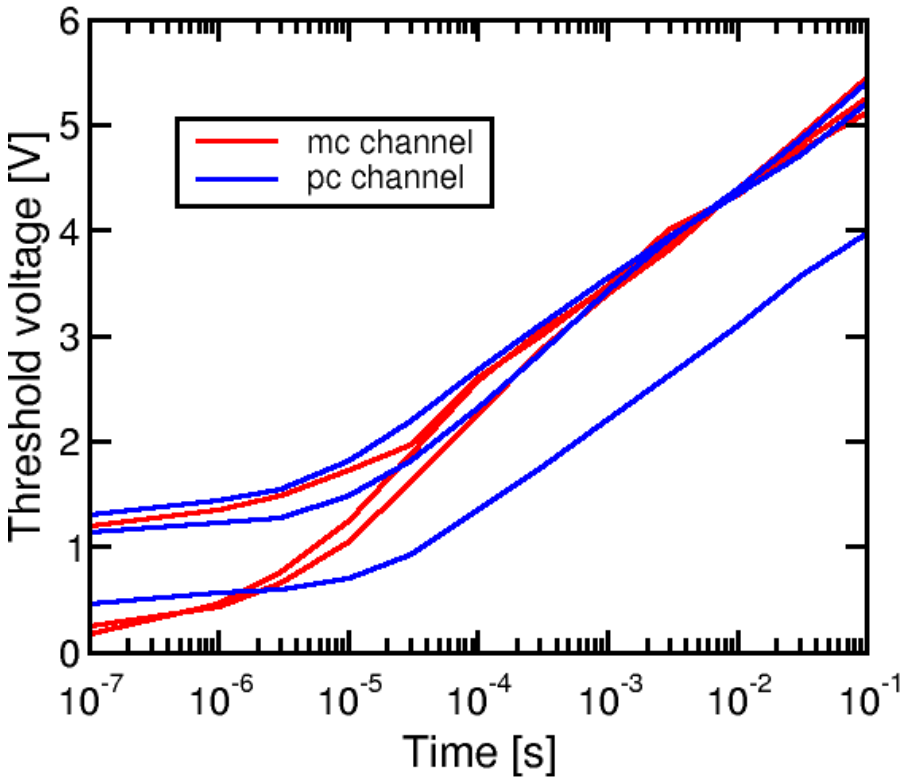


Figura 4.2.6: confronto di curve di programmazione tra celle TANOS floating body con substrato policristallino e monocristallino (tensione di comando 18V).

4.2.1 Meccanismi di programmazione

Il meccanismo di programmazione della stringa TANOS floating body è stato esplorato in maniera dettagliata. Grazie ad alcuni accorgimenti applicati in fase di misura è stato possibile polarizzare in modo non convenzionale i terminali della stringa applicando anche tensioni negative. Lo scopo è quello di capire con che velocità il substrato reagisce all'impulso applicato e con quale costante di tempo risponde alla polarizzazione di BL e SL in relazione alla capacità associata. Questa valutazione può essere fatta applicando una tensione diversa da zero a BL e SL e vedere con che velocità il substrato si polarizza, mantenendo la stessa differenza di potenziale ai capi dell'ossido. Bisogna quindi osservare il comportamento delle curve di programmazione in differenti condizioni di applicazione dell'impulso di programmazione.

Le tre celle qui presentate e sperimentalmente osservate appartengono ad un wafer TANOS floating body con substrato in silicio policristallino e con impianti di giunzione tiltati. E' stata dapprima ricavata una curva di programmazione con polarizzazione standard, successivamente è stato applicato un impulso di cancellazione che riportasse la tensione di soglia vicino al valore vergine. In una programmazione standard, ai capi della cella selezionata viene generato un campo elettrico applicando 18V alla WL e mantenendo la BL a massa. L'idea è stata quella di mantenere inalterata la differenza di potenziale generatrice del campo traslando verso il basso le tensioni applicate. In un primo caso la BL è stata polarizzata a -18V mentre la WL selezionata è stata tenuta a massa; anche gli altri terminali sono stati polarizzati per mantenere ai loro capi le differenze di potenziale del caso standard, applicando quindi -15V ai selettori DSL e SSL e -10V alle WL non selezionate. In un secondo caso sono state applicate le tensioni negative ai vari terminali mentre la WL selezionata è stata mantenuta flottante. In un terzo caso, sempre a tensioni negative, tutte le WL della stringa sono state mantenute flottanti, leggendo sempre una sola cella d'interesse. L'operazione di lettura non è stata alterata. I risultati della misura sono presentati in figura 4.2.7.

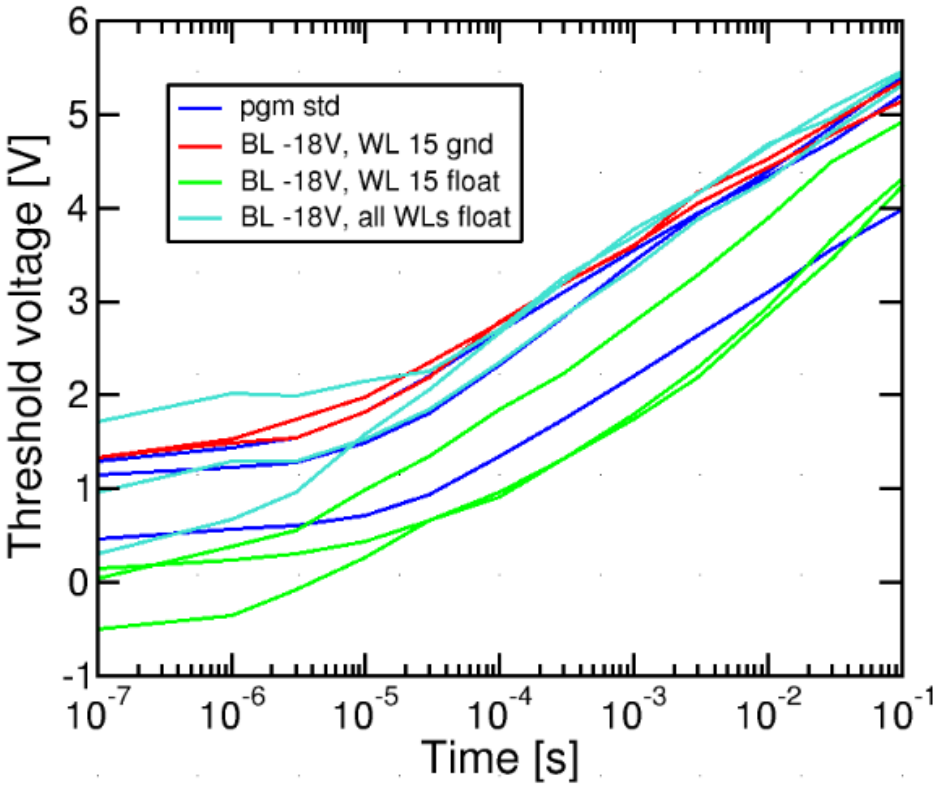


Figura 4.2.7: curve di programmazione a diverse condizioni di polarizzazione di celle TANOS floating body con substrato in silicio policristallino e impianti di giunzione tiltati.

Come si può vedere, le curve mantengono in tutti i casi andamenti simili. Questo comportamento mette in luce le caratteristiche di simmetria della stringa flottante rispetto alle diverse modalità di programmazione. Non permette tuttavia di ricavare informazioni sulla costante di tempo con cui il substrato reagisce ad un segnale di polarizzazione di BL e SL poiché l'impulso programma la cella con lo stesso trend del caso standard, anche per tempi brevi.

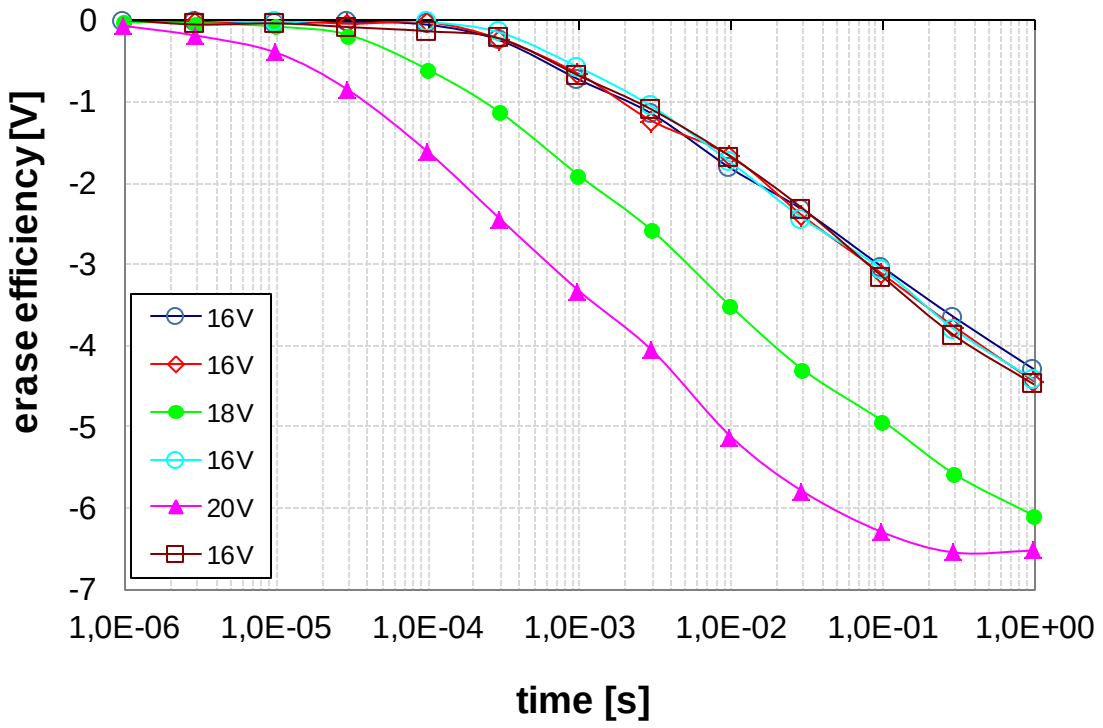
4.3 Curve di cancellazione

Come visto nei precedenti capitoli, l'operazione di cancellazione delle stringhe NAND implica l'applicazione di una tensione al contatto di substrato in modo che il campo elettrico che si genera ai capi della cella permetta l'emissione dal nitruro di trapping delle cariche elettriche verso il substrato stesso. Nelle celle TANOS a body flottante non è possibile eseguire questa operazione poiché non è possibile contattare

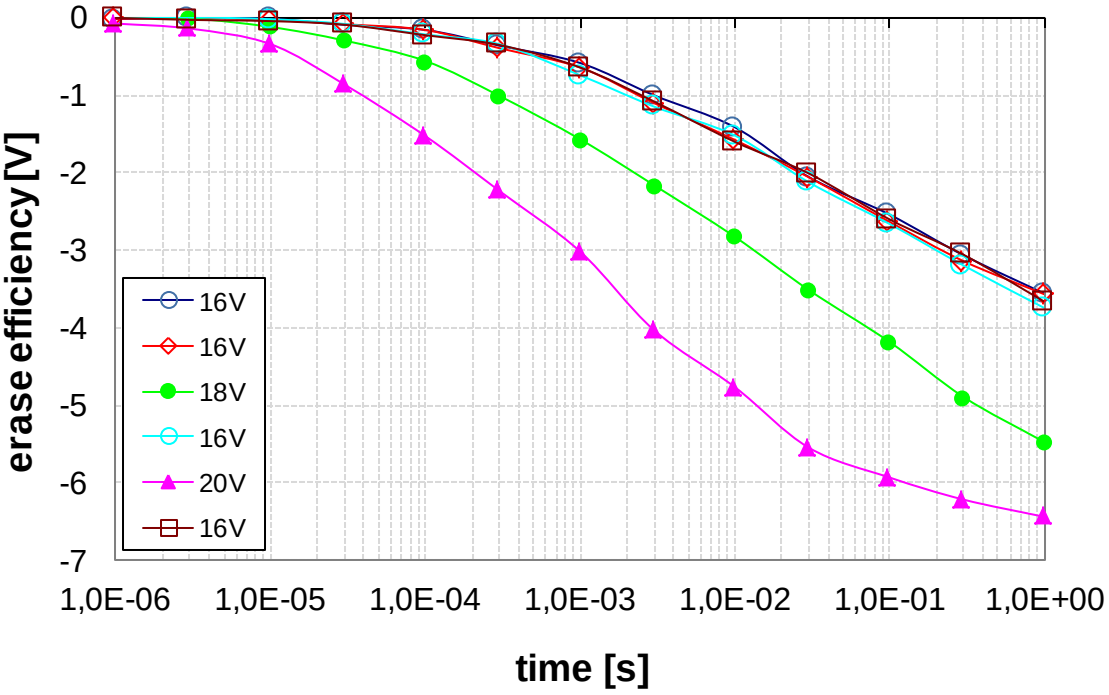
direttamente il substrato. Per riuscire a generare la corretta differenza di potenziale tra substrato e gate onde cancellare la cella si sfruttano le giunzioni laterali n^+ - substrato p. Queste giunzioni sono a tutti gli effetti dei diodi per i quali, se viene applicata al catodo (zona n^+) una tensione positiva rispetto all'anodo (substrato p), viene forzata una corrente inversa ovvero un moto di cariche attraverso la giunzione verso il substrato. Le cariche dalla zona n^+ raggiungono il substrato p e causano un innalzamento del potenziale di substrato, pertanto mantenendo le WL a massa è possibile creare il campo elettrico necessario per estrarre le cariche dal nitruro. Nelle operazioni di cancellazione il substrato viene quindi contattato indirettamente per mezzo di una polarizzazione applicata ai terminali di BL e SL e l'efficacia con cui si modifica il potenziale di substrato dipende dalle giunzioni e dalla loro corrente inversa.

La polarizzazione della stringa in fase di cancellazione è la seguente: le WL sono mantenute a massa, i selettori vengono portati a 8V (tensione determinata per via sperimentale per la quale si minimizza il disturbo indotto dall'operazione di cancellazione sui selettori stessi, i dettagli verranno descritti in seguito) e BL e SL vengono portati alla tensione di comando. Dalla figura 4.3.1 si nota che le celle TANOS a body flottante con o senza giunzioni possono essere cancellate sfruttando il meccanismo sopra descritto; inoltre presentano una finestra ampia e un andamento della tensione di soglia confrontabile di quello della tecnologia a body contattabile (figura 4.3.2).

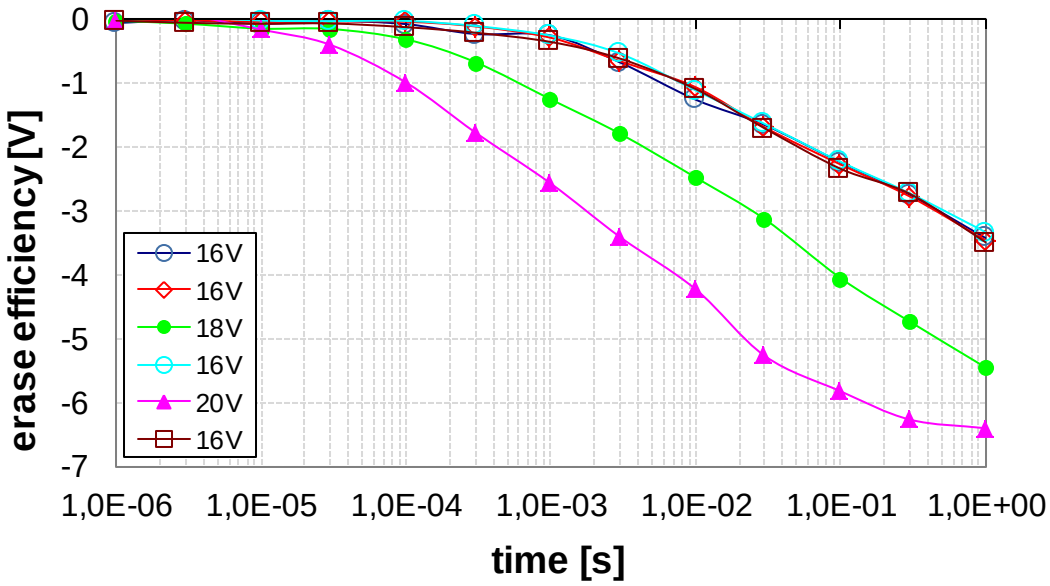
Nella figura 4.3.3 sono presentate le curve di efficienza di cancellazione di stringhe con substrato monocristallino, per le quali vengono confermati i benefici relativi alla maggiore corrente discussi nel precedente capitolo dovuti alla minore difettosità del silicio del body. Un parallelo tra le curve relative alle differenti composizioni cristallografiche dei substrati è presente in figura 4.3.4, in cui si evidenziano andamenti della soglia sostanzialmente confrontabili.


0
-1
-2
-3
-4
-5
-6
-7
erase efficiency [V]
16V
16V
18V
16V
20V
16V
1,0E-06
1,0E-05
1,0E-04
1,0E-03
1,0E-02
1,0E-01
1,0E+00
time [s]

(a)


0
-1
-2
-3
-4
-5
-6
-7
erase efficiency [V]
16V
16V
18V
16V
20V
16V
1,0E-06
1,0E-05
1,0E-04
1,0E-03
1,0E-02
1,0E-01
1,0E+00
time [s]

(b)



(c)

Figura 4.3.1: curve di efficienza di cancellazione di celle TANOS a body flottante con substrato in silicio policristallino e impianti di giunzione (a), impianti di giunzione mascherati (b), impianti di giunzione tiltati (c). Le curve hanno all'incirca gli stessi andamenti e raggiungono gli stessi valori di efficienza di cancellazione.

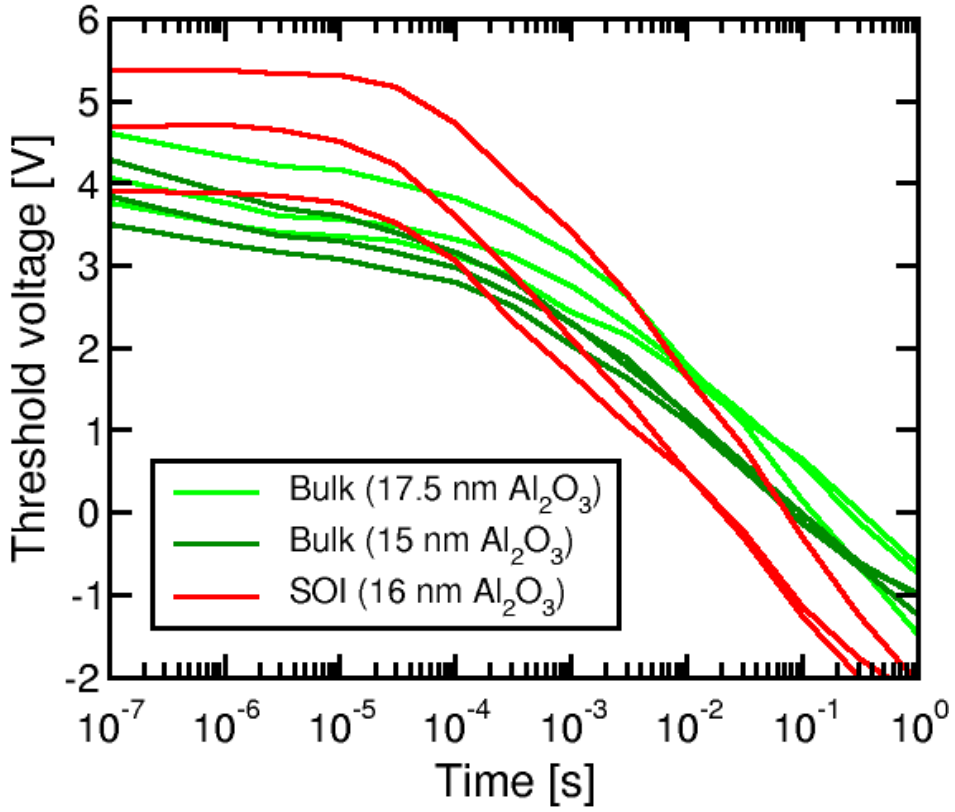


Figura 4.3.2: confronto di efficienze di cancellazione tra celle TANOS standard e TANOS floating body (tensione di comando 18V).

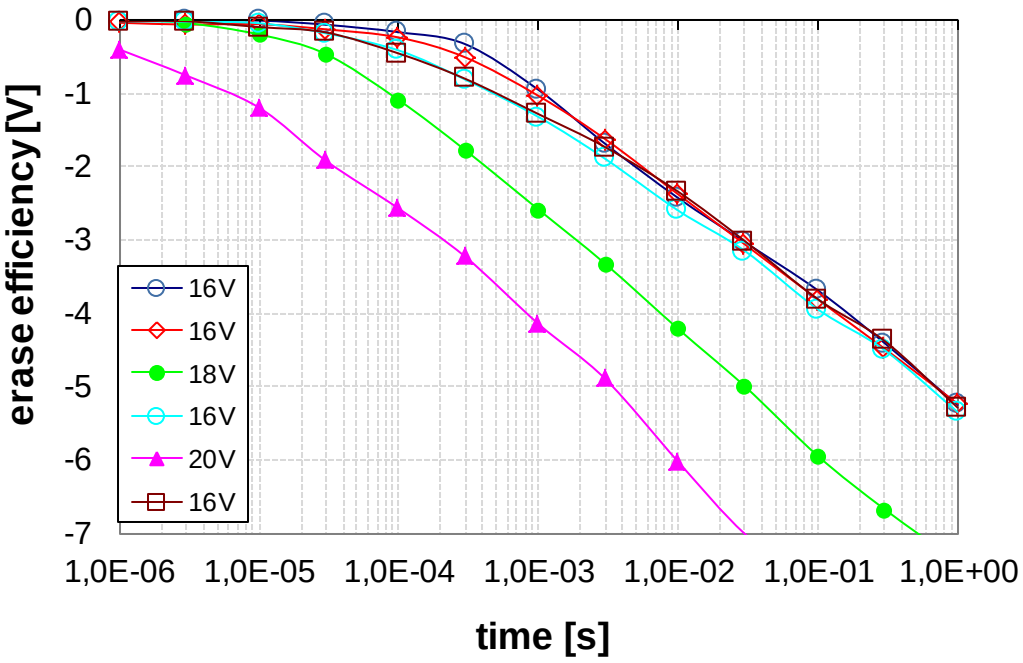


Figura 4.3.3: curve di efficienza di cancellazione di celle TANOS a body flottante con substrato in silicio monocristallino e impianti di giunzione tiltati.

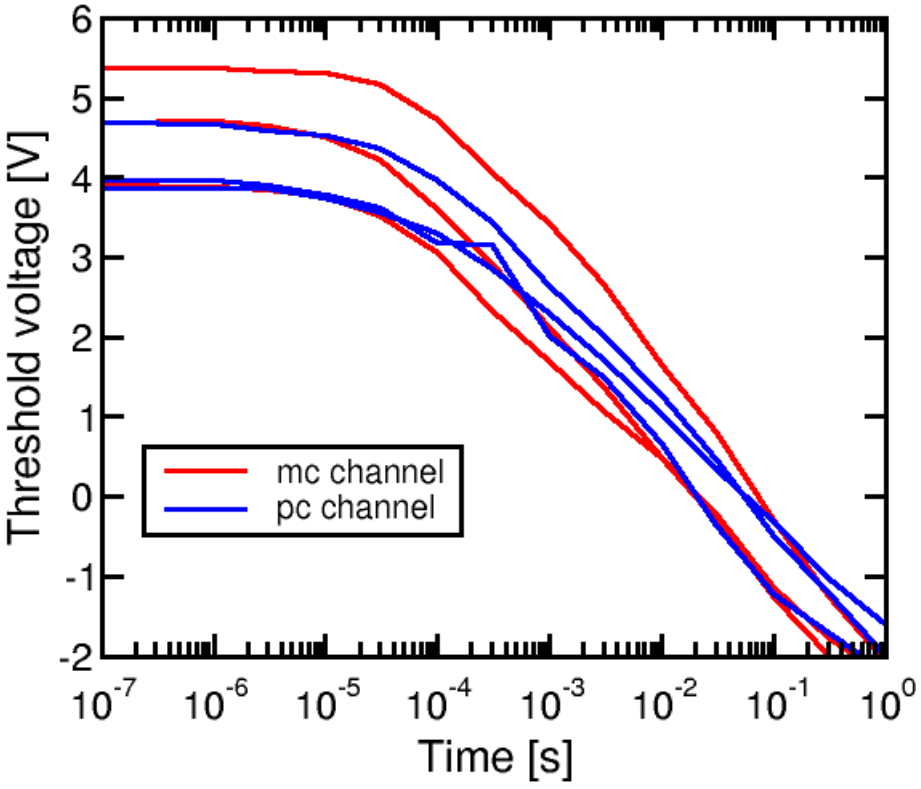


Figura 4.3.4: confronto di curve di cancellazione tra celle TANOS a substrato policristallino e monocristallino (tensione di comando 18V).

4.3.1 Comportamento dei selettori

Sono state effettuate apposite misure per monitorare l'andamento delle tensioni di soglia dei selettori in funzione della polarizzazione loro applicata in fase di cancellazione. Il comportamento dei selettori e della cella sono illustrati nella figura 4.3.3. La misura è stata effettuata ripetendo programmazioni e cancellazioni a tempo costante variando solamente la tensione di selettore: è stato esplorato il range di tensione da 1V a 18V ed è stato osservato che non vi è variazione di tensione di soglia del selettore per un valore di tensione applicata di 8V.

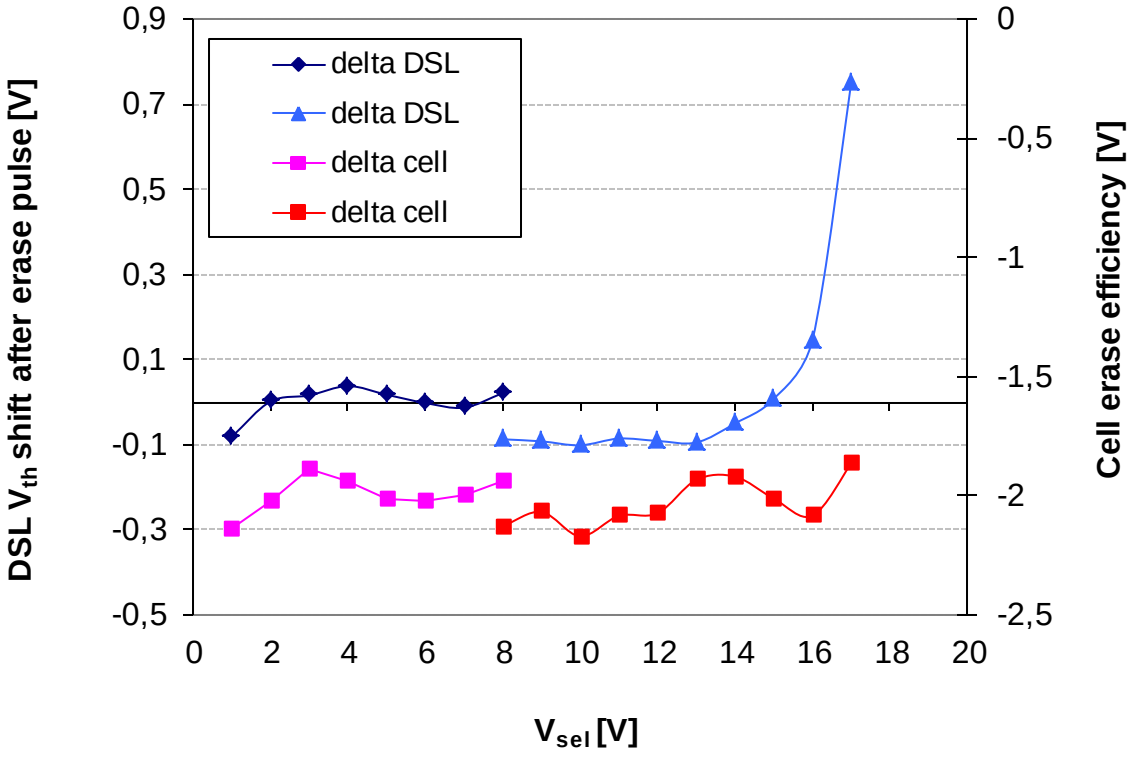


Figura 4.3.3: scostamento delle soglie del selettore DSL e della cella al variare della polarizzazione del selettore in fase di cancellazione

Se viene scelta una tensione di gate del selettore troppo elevata, l'evidenza sperimentale mostra uno scostamento positivo della soglia del selettore, che quindi modifica il suo stato verso quello programmato. Questo effetto, come accennato nel precedente capitolo, è dovuto al fatto che i selettori sono realizzati in modo molto simile alle altre celle, con uno strato di nitruro di trapping nello stack del gate, perciò si comportano come celle di memoria e vanno correttamente polarizzati per evitare

scostamenti di soglia: perciò se al selettore viene applicata una tensione alta, maggiore di 12V, vede ai suoi capi una tensione tale da favorire la programmazione. Viceversa, se la tensione di polarizzazione è troppo bassa, lo scostamento di soglia è negativo ovvero il selettore tende a cancellarsi. Bisogna quindi scegliere una tensione tale per cui i selettori non modifichino le loro tensioni di soglia durante le operazioni di cancellazione: per questo motivo DSL e SSL sono polarizzati a 8V.
E' interessante notare che l'efficienza di cancellazione della cella non risente della polarizzazione applicata ai selettori.

4.3.2 Cancellazione "single side"

Per come è stato precedentemente descritto, il meccanismo di cancellazione nelle stringhe a body flottante avviene per mezzo dell'iniezione di cariche attraverso le giunzioni laterali connesse ai terminali di BL e SL che polarizzano indirettamente il substrato. In linea teorica e concettuale, ci si può aspettare un effetto di cancellazione anche applicando la tensione di comando da un solo lato (BL o SL). Per questo si è sperimentata la cancellazione "single side": iniettare le cariche solamente da uno dei due lati e forzare la corrente inversa solo attraverso una giunzione. La tensione di cancellazione di 18V è stata applicata solamente al contatto di BL (oppure al contatto di SL) polarizzando il selettore DSL a 8V mentre dall'altro lato la SL è stata lasciata flottante con il selettore SSL a massa. I risultati sono descritti dalla figura 4.3.4.

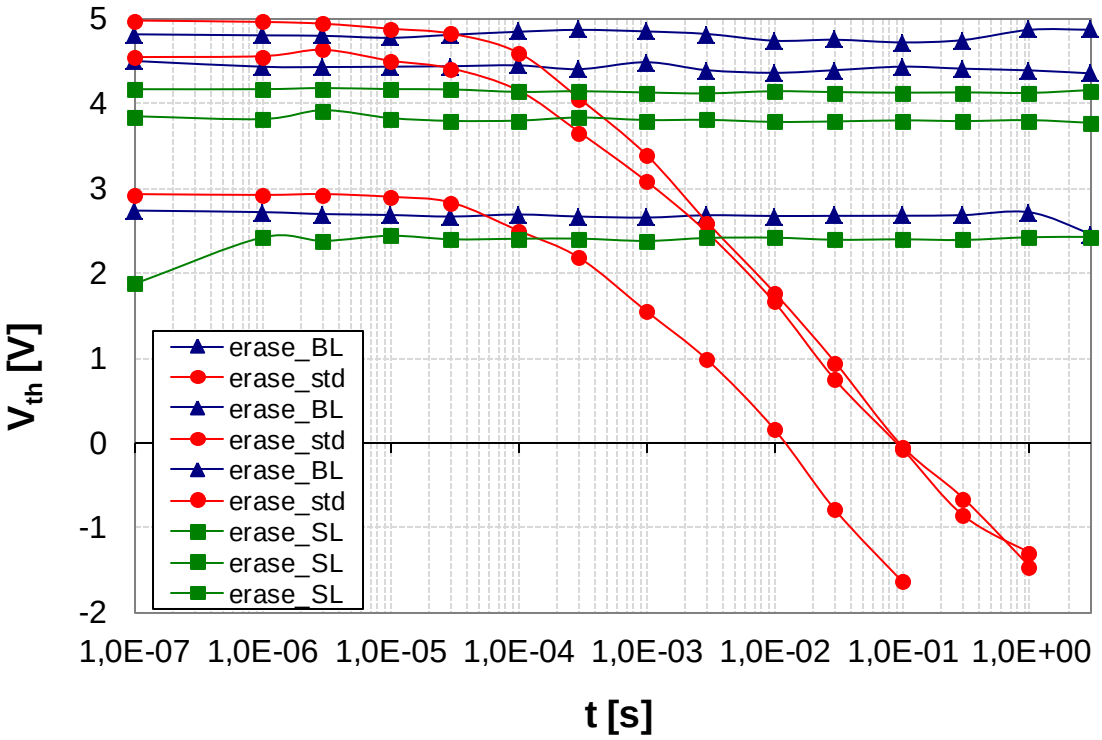


Figura 4.3.4: curve di cancellazione di celle TANOS a body flottante con substrato in silicio policristallino e impianti di giunzione tiltati. Le curve piatte sono relative alle misure di cancellazione dal solo lato BL oppure SL.

Il risultato sperimentale evidenzia come la cancellazione da un singolo lato non sia efficace anche per elevati tempi di applicazione della tensione di comando di 18V alla BL. Per cercare di renderla efficace è stata applicata una differente polarizzazione al SSL, senza tuttavia giungere a risultati soddisfacenti. Lo stesso tipo di procedura è stata utilizzata per misurare l'efficienza di cancellazione "single side" di celle di memoria con substrato in silicio monocristallino senza tuttavia osservare alcuna variazione della tensione di soglia al variare della durata dell'impulso di comando applicato: per le misure condotte, non risulta possibile cancellare la cella polarizzando un solo lato. Questo risultato è dovuto al fatto che in queste misure il substrato non viene polarizzato alla tensione di comando imposta per esempio sulla BL, ma rimane fisso alla tensione più bassa relativa all'altro terminale SL della stringa, rendendo inefficace la cancellazione.

4.3.3 Meccanismi di cancellazione

Come nel caso della programmazione, il meccanismo di cancellazione è stato esplorato in maniera accurata tramite misure sperimentali che hanno sfruttato alcuni particolari accorgimenti. Si è voluto verificare principalmente la simmetria dell'operazione di cancellazione e ricavare le capacità associate ai terminali cui viene applicata la tensione di comando (WL), traslando le tensioni verso il basso e mantenendo costanti le differenze di potenziale che si sviluppano ai capi delle celle. I confronti tra gli andamenti delle curve di cancellazione sono rappresentati nella figura 4.3.5 in cui sono descritte le curve relative a tre siti cui sono state applicate differenti condizioni di polarizzazione, comparate con quelle frutto del meccanismo standard per la cancellazione a body flottante precedentemente descritta. Prima di ogni nuova operazione di cancellazione le celle sono state riprogrammate in modo da raggiungere un valore di soglia che potesse far apprezzare lo scostamento tra gli stati programmato e cancellato.

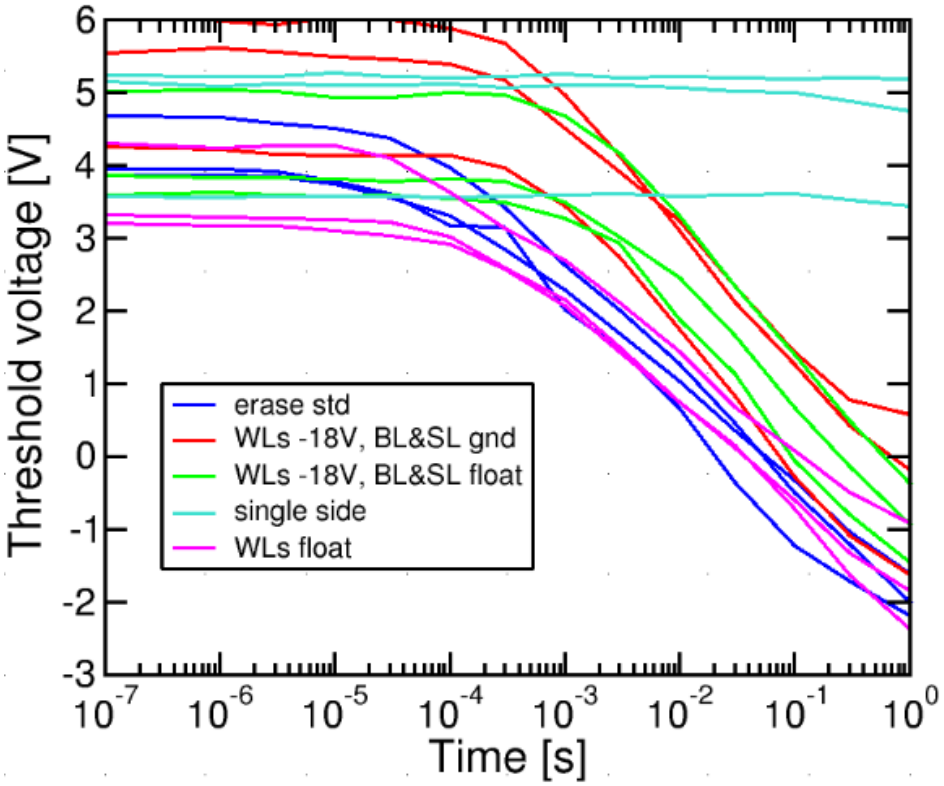


Figura 4.3.5: curve di cancellazione a diverse condizioni di polarizzazione di celle TANOS floating body con substrato in silicio policristallino e impianti di giunzione tiltati.

Come si può notare, il meccanismo risulta simmetrico rispetto alle tensioni applicate: la cella viene efficacemente cancellata applicando tensioni negative a BL e SL e mantenendo tutte le WL a massa oppure flottanti, in modo da generare ai capi delle celle della stringa la differenza di potenziale necessaria per la cancellazione. Nemmeno con queste misure si è riuscito a ricavare il peso capacitivo dei terminali ai quali viene applicata la tensione di comando, poiché le curve sono tutte allineate.
E' stata applicata anche una polarizzazione simile a quella del caso di cancellazione standard con BL e SL a 18V mentre tutte le WL sono state lasciate flottanti, osservando che anche in queste condizioni l'andamento delle curve segue quello del caso di cancellazione standard. Infine, sono presenti anche le curve delle celle che sono state cancellate polarizzando un solo lato della stringa (per i dettagli si veda il precedente paragrafo).

4.3.4 Andamento del potenziale nella stringa

L'operazione di cancellazione per le architetture TANOS floating body, così come ogni altra architettura NAND, agisce su interi settori e non sulle singole celle. Volendo analizzare ulteriormente il meccanismo di cancellazione di stringhe di celle a body flottante si è cercato di capire se l'effetto di cancellazione fosse più o meno efficace per le celle più vicine zona di contatto di BL e alle giunzioni attraverso cui passa la corrente inversa che polarizza il substrato.

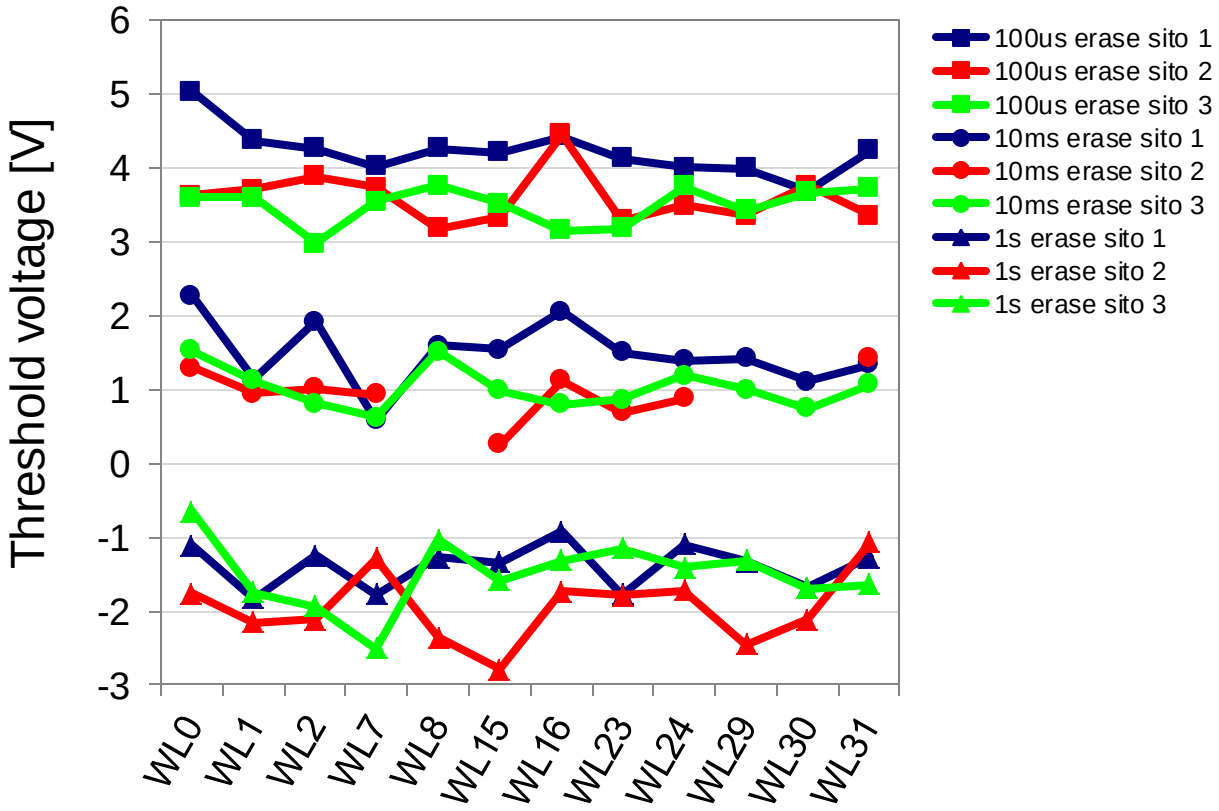


Figura 4.3.6: andamento del profilo di potenziale all'interno di una stringa TANOS floating body con impianti tiltati.

Per questo sono state condotte misure di cancellazione andando a monitorare il comportamento delle tensioni di soglia delle differenti WL in relazione alla durata del tempo di applicazione del segnale di comando; è stato quindi ricostruito l'andamento delle tensioni di soglia all'interno della stringa per ogni WL misurata (fig. 4.3.6). Come si può vedere l'operazione di cancellazione coinvolge le varie WL senza influenzarne maggiormente nessuna in particolare: la cancellazione è efficace in modo pressoché costante lungo tutta la stringa, senza preferire le WL più vicine al punto in cui è applicata la tensione.

4.4 Ciclatura

Come abbiamo visto, le operazioni di programmazione e cancellazione sulle celle TANOS a body flottante risultano efficaci. Riprendendo i concetti introdotti nei capitoli precedenti, sappiamo che la ripetizione di queste operazioni è causa di un

degrado intrinseco nell'utilizzo della cella stessa che può essere valutato dalle misure di ciclatura.

Gli esperimenti condotti hanno visto la programmazione e la cancellazione (ciclatura) di un certo numero di campioni per ottenere una buona statistica dei risultati. Le misure sono state effettuate misurando ad intervalli logaritmici i valori di tensione di soglia ottenuti dopo un certo numero di cicli fino ad arrivare a circa 27000 cicli. Le programmazioni sono state effettuate con tensione di comando di 18V per la durata di 300us; le cancellazioni hanno avuto la stessa tensione di comando e durata degli impulsi di 10ms. Non è stato possibile applicare un treno di impulsi (come per le misure delle celle TANOS a body contattabile) e misurare le soglie dopo un certo numero di cicli per problemi di setup sperimentale, tuttavia le misure sono state effettuate in modo che i dati fossero perfettamente confrontabili.

Durante le fasi di programmazione e cancellazione della cella si ha un flusso di cariche (elettroni o lacune) attraverso l'ossido di bottom. Il passaggio di questa corrente di tunnel è un processo distruttivo che genera difetti e stati interfacciali responsabili del degrado durante la ciclatura. In particolare, nelle celle TANOS durante la cancellazione si ha un flusso sia di elettroni che di lacune: l'influenza dei difetti generati da questi due flussi è tale per cui lo scarto tra due livelli di soglia risulta simmetrico.

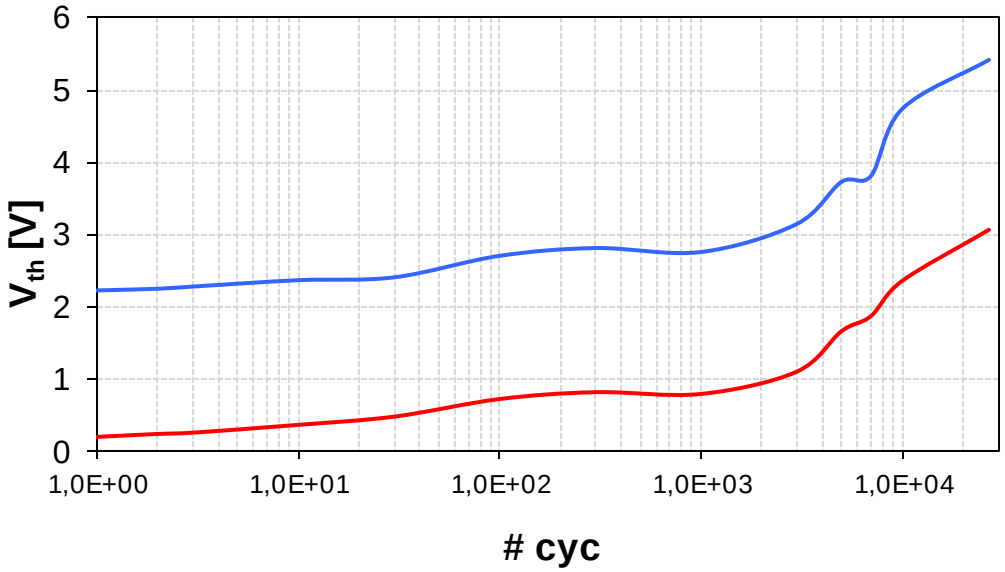
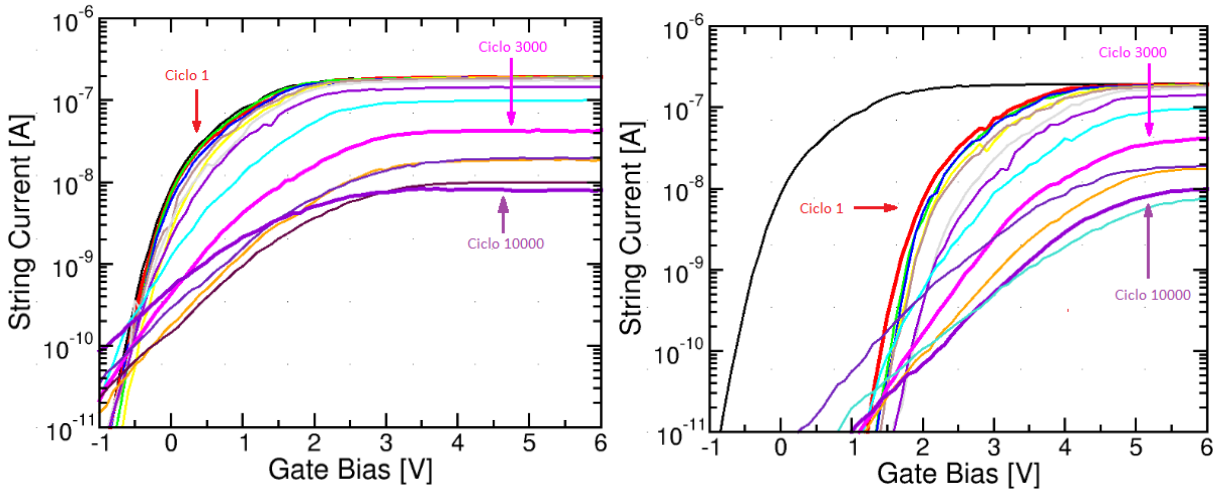


Figura 4.4.1: finestra di ciclatura per celle di memoria TANOS floating body con substrato in silicio policristallino e impianti di giunzione tiltati. La curva in blu mostra il degrado in fase di programmazione, l'altra in fase di cancellazione.

La figura 4.4.1 mostra i dati sperimentali come il risultato di una media dei valori di soglia ottenuti che valutano il degrado della tensione di soglia in fase di programmazione e in fase di cancellazione. All'interno di questa statistica, si può notare che le due curve hanno sostanzialmente lo stesso andamento, inoltre la soglia dello stato programmato e quella dello stato cancellato sono distinguibili fino a circa 10000 cicli, successivamente il degrado in cancellazione porta la soglia cancellata a valori troppo elevati per poter garantire il corretto funzionamento del dispositivo.

Di seguito, in figura 4.4.2, vengono presentati gli andamenti delle transcaratteristiche di due campioni misurati in ciclatura. Come si può notare, al ripetersi delle operazioni di programmazione e di cancellazione, si osservano due differenti effetti: dapprima una traslazione rigida della transcaratteristica, mentre dopo un certo numero di cicli si osserva un piegamento delle curve con corrispondente aumento della pendenza di sottosoglia e diminuzione della corrente di saturazione che scorre nel transistor. Questo effetto è dovuto principalmente alla presenza di stati-trappola interfacciali che si generano con il flusso di cariche attraverso l'ossido: all'aumento della densità di

difetti superficiali si ha un corrispondente aumento della capacità ad essi associata, con conseguente degrado della pendenza sottosoglia [17].



(a) (b)

Figura 4.4.2: andamenti delle transcaratteristiche durante la ciclatura in fase di cancellazione (a) e programmazione (b). Si nota il degrado di pendenza e corrente di saturazione all'aumentare con il numero di cicli.

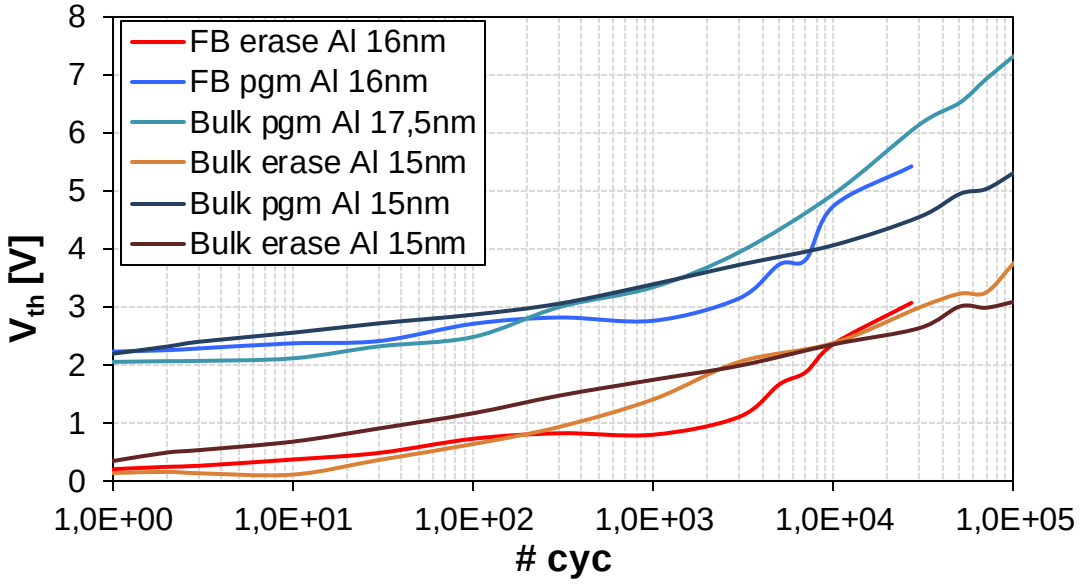


Figura 4.4.3: confronto tra finestre di ciclatura per celle di memoria TANOS floating body (FB) e TANOS standard planari (Bulk).

In figura 4.4.3 è presente un confronto dei dati di degrado in ciclatura tra tecnologia a substrato flottante e tecnologia a substrato contattabile: si può notare come le prestazioni tra le due tecnologie siano allineate all'incirca fino ai 10000 cicli.

4.5 Ritenzione

Come è stato accennato anche nei precedenti capitoli, per queste tecnologie le prestazioni di ritenzione sono uno dei parametri critici fondamentali per valutare l'affidabilità e la qualità di una memoria non volatile.

Per valutare la ritenzione, ovvero la capacità di una cella di mantenere inalterato il suo stato e conservare il dato per lunghi periodi di tempo, sono state effettuate diverse misure. Alcune di queste sono state compiute a temperatura ambiente, programmando diverse celle TANOS a body flottante con una tensione di comando di 18V per una durata dell'impulso variabile (10ms oppure 1ms) e andando a leggere ad intervalli di tempo logaritmici la tensione di soglia per poi fare una media di tutti i campioni misurati, monitorando le variazioni della tensione rispetto alla soglia vergine. I risultati di questo metodo sono dipinti in figura 4.5.1: si può notare che dopo circa 4000 minuti (corrispondenti a circa 66 ore) la perdita in ritenzione si aggira intorno agli 0,3V, un valore confrontabile con quelli delle celle SONOS a body contattabile e migliore di quello a cui si fa riferimento nel capitolo precedente per quanto riguarda le TANOS a body contattabile.

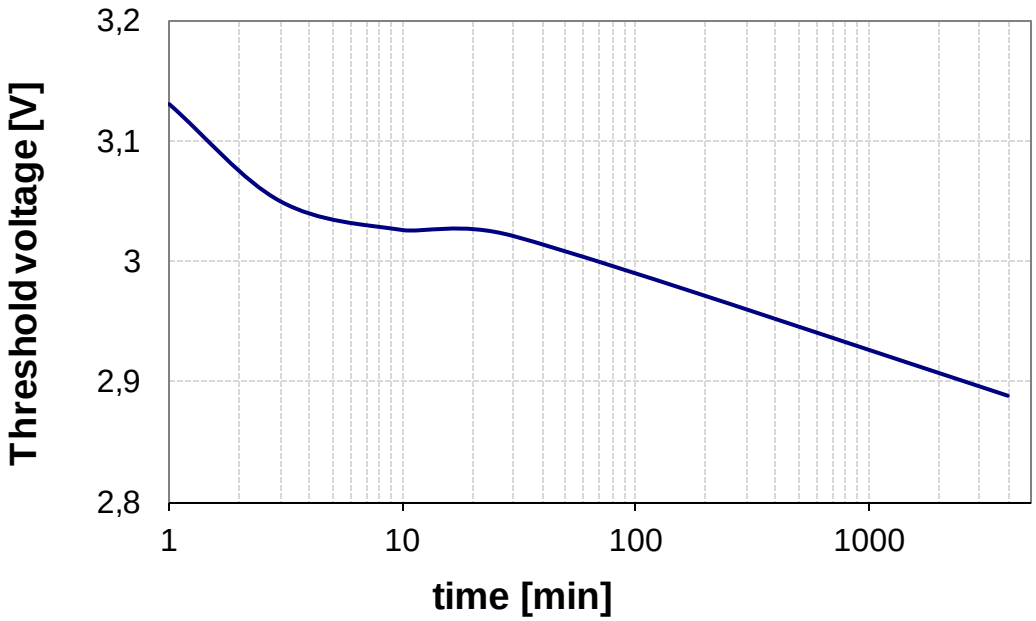


Figura 4.5.1: andamento della tensione di soglia nelle celle TANOS a body flottante

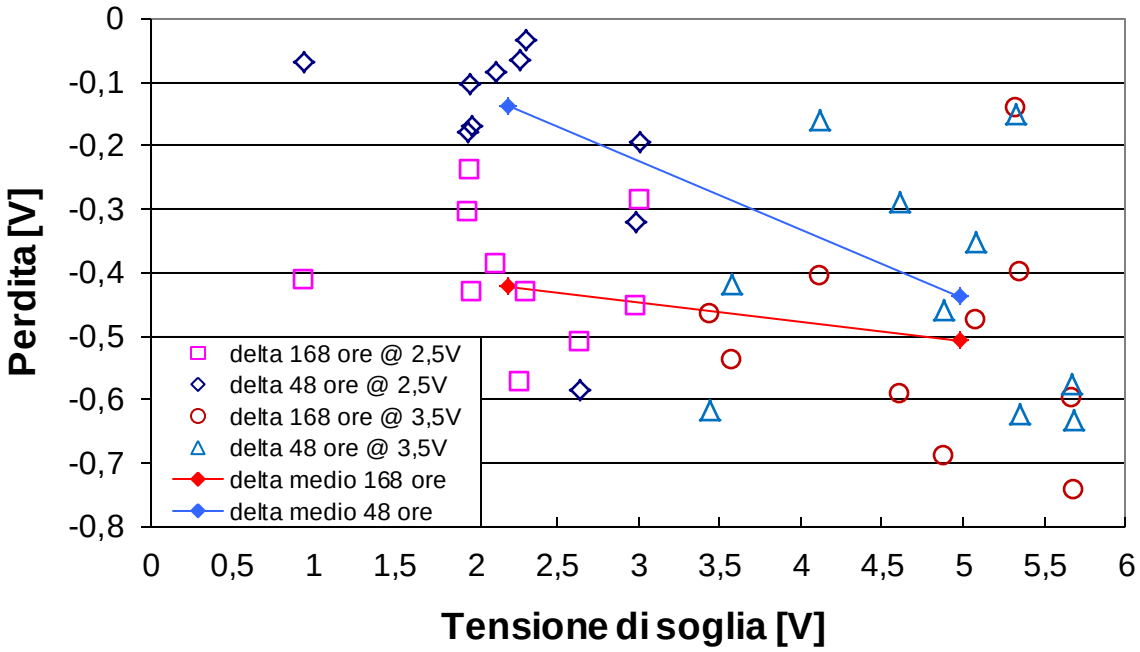


Figura 4.5.2: estrazione del valore di perdita in ritenzione per celle di memoria TANOS floating body a substrato policristallino con tensione di soglia di 3,5V.

Un'altra metodologia è stata applicata per estrapolare il medio valore della perdita di tensione come differenza tra la soglia programmata e la soglia vergine. Alcune celle dello stesso wafer sono state suddivise in due gruppi, programmati entrambi con una tensione di 18V ma con tempi differenti, il primo con impulsi da 1ms, il secondo con

impulsi da 10ms. Il wafer è stato successivamente inserito in un forno (bake) a 60° per stimolare l'emissione termoionica degli elettroni dallo strato di trapping ed è stata misurata la tensione di soglia delle celle dopo 48 ore e dopo 168 ore (figura 4.5.2). Ricavando il valore medio della tensione di soglia ne è emerso che per celle programmate con soglia di 3,5V la perdita media dopo 48 ore è di poco inferiore a 0,3V mentre si alza a circa 0,47V alle 168 ore. Dalle evidenze sperimentali, è lecito pensare che le eventuali variazioni delle prestazioni in ritenzione non dipendano dalla presenza del substrato flottante.

4.6 Conclusioni

In questo capitolo abbiamo presentato le strutture che uniscono i benefici dell'idea del charge trapping con la possibilità di creare celle di memoria di tipo SOI, realizzando le TANOS floating body. Sono state analizzate le principali caratteristiche della cella così realizzata confrontandole con le strutture a body contattabile descritte nel capitolo precedente: le prestazioni delle tecnologie SOI sono allineate con quelle a body contattabile eccezion fatta per i livelli di corrente che sono diminuiti di poco meno di un ordine di grandezza. Soprattutto è stata messa in risalto la possibilità di cancellazione per la cella a body flottante senza dover contattare direttamente il substrato, oltre che l'efficacia delle strutture junctionless non differenti dal punto di vista delle prestazioni dalle classiche architetture con giunzioni interne. Le buone qualità delle celle TANOS a body flottante sono da tenere in considerazione per la realizzazione di celle di memorie stacked e 3D.

Nel prossimo capitolo vedremo un'altra tipologia di cella di memoria SOI, chiamata SONOS floating body, e ne confronteremo le caratteristiche con le celle finora analizzate.

Capitolo 5

Celle di memoria SONOS FinFET

5.1 Struttura

Le architetture delle celle di memoria SONOS sono state introdotte nel capitolo 2 in cui sono stati presentati pregi e difetti di questo tipo di dispositivi. In particolare, è stata approfondito un effetto relativo all'operazione di cancellazione, l'*erase saturation*. L'iniezione di carica tra il gate in polisilicio verso il nitruro attraverso il sottile ossido di top, dovuta al campo elettrico applicato necessario per cancellare, impone un limite all'efficienza di estrazione di carica dallo strato di trapping verso il substrato: quando i due flussi di carica si bilanciano, la tensione di soglia della cella raggiunge un valore di saturazione al di sotto del quale non è possibile scendere. Per questo motivo, le celle SONOS a body flottante sono state realizzate per mezzo di un etching anisotropo in modo che la loro area attiva (substrato di silicio monocristallino) abbia una forma arrotondata; analogamente gli strati di ossido di bottom (diossido di silicio SiO_2), nitruro (Si_3N_4) e ossido di top (ancora diossido di silicio SiO_2) cresciuti sul substrato seguono lo stesso profilo. Applicando gli impulsi di cancellazione, le linee del campo elettrico nel caso planare sono solamente verticali attraverso una superficie limitata, mentre nel caso di SONOS con area attiva arrotondata il flusso del campo elettrico è molto maggiore per l'aumento della superficie dell'ossido di bottom che segue il profilo arrotondato del substrato (figura 5.1.1 e figura 5.1.2). L'applicazione di una polarizzazione aumenta dunque il campo elettrico efficace: si cerca di superare il limite imposto dall'*erase saturation* per la cancellazione della cella, andando ad agire con maggior forza sulle cariche intrappolate semplicemente ingegnerizzando la morfologia del transistor senza modificare le tensioni di comando o altri parametri

elettrici. Le celle SONOS così realizzate sono molto simili ai FinFET descritti nel capitolo 2.

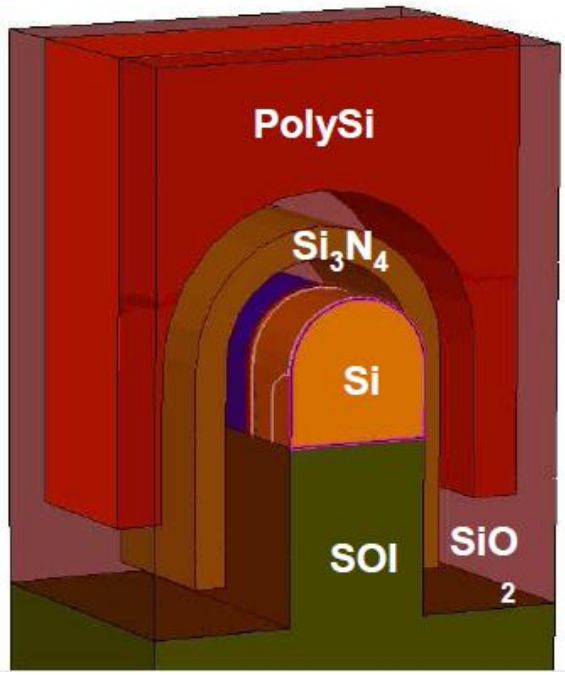


Figura 5.1.1: schema dell'architettura di una cella SONOS floating body con area attiva tonda.

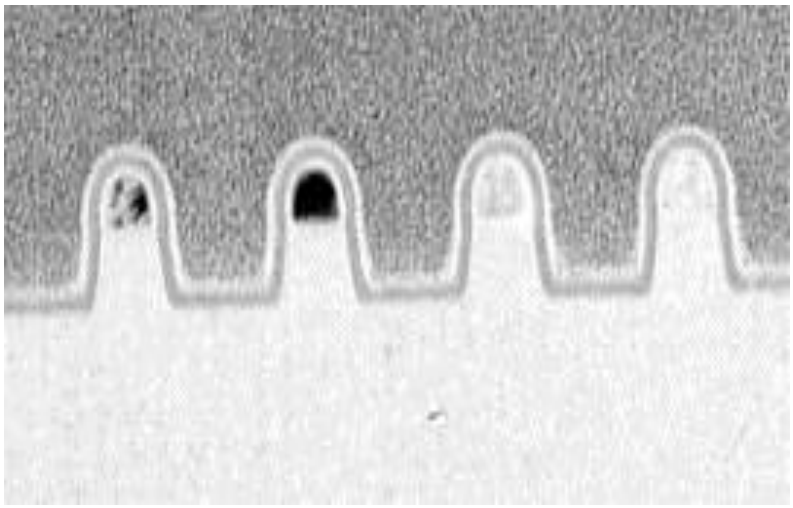


Figura 5.1.2: immagine TEM della struttura della stringa SONOS floating body con area attiva tonda.

Le celle SONOS oggetto di misura sono realizzate con spessori nominali dello stack del gate pari rispettivamente a 5nm per l'ossido di top, 6nm per il nitruro di trapping e 4,5nm per l'ossido di bottom. Nella tabella 5.1.3 sono riportate le caratteristiche tecnologiche delle celle presentate in questo capitolo.

substrato	bottom oxide	nitruro	top oxide
Si p-type 25nm	SiO_2 4,5nm÷4nm	SiN LPCVD 6nm	SiO_2 5nm

Figura 5.1.3: caratteristiche e spessori nominali

Le misure sono state effettuate tramite probe station contattando i terminali delle stringhe di transistors realizzati su wafer di silicio. Ogni cella è identificata da un contatto di Word Line da 0 a 31, al terminale della cella selezionata è stato applicato un impulso di tensione in fase di programmazione o di cancellazione, misurando successivamente la corrente per ricavarne la transcaratteristica ed estrarre la tensione di soglia corrispondente ad un fissato valore di corrente (10nA). Le tensioni di comando nominali sono state 16V e 18V, poiché la delicatezza della struttura non ha permesso l'applicazione di stress elettrici maggiori, mentre i tempi per i quali è stato applicato l'impulso sono cumulativi in scala logaritmica da 1us a 100ms (1s per la cancellazione).

5.2 Curve di programmazione

Dal punto di vista della programmazione, la stringa è stata polarizzata come segue: la tensione di comando è stata applicata alla WL selezionata mentre tutte le altre WL sono state portate alla tensione V_{pass} pari a 8V, tale da permettere l'inversione del substrato senza inficiare programmazioni indesiderate; i selettori DSL e SSL sono stati portati a 3V, la BL è stata portata a 0V e la SL a 3V (in modo esattamente identico a quanto fatto per le celle TANOS, e quindi ad una stringa NAND). Successivamente si è

andati a leggere transcaratteristica estrapolandone la tensione di soglia che rappresenta lo stato programmato della cella. E' da notare come in fase di programmazione ed in fase di lettura, esattamente come descritto precedentemente nella caratterizzazione delle celle TANOS floating body e delle TANOS planari standard, il substrato non venga polarizzato ma venga mantenuto flottante. Il grafico riportato in figura 5.2.1 presenta un confronto delle efficienze in dipendenza dai diversi spessori di ossido di bottom.

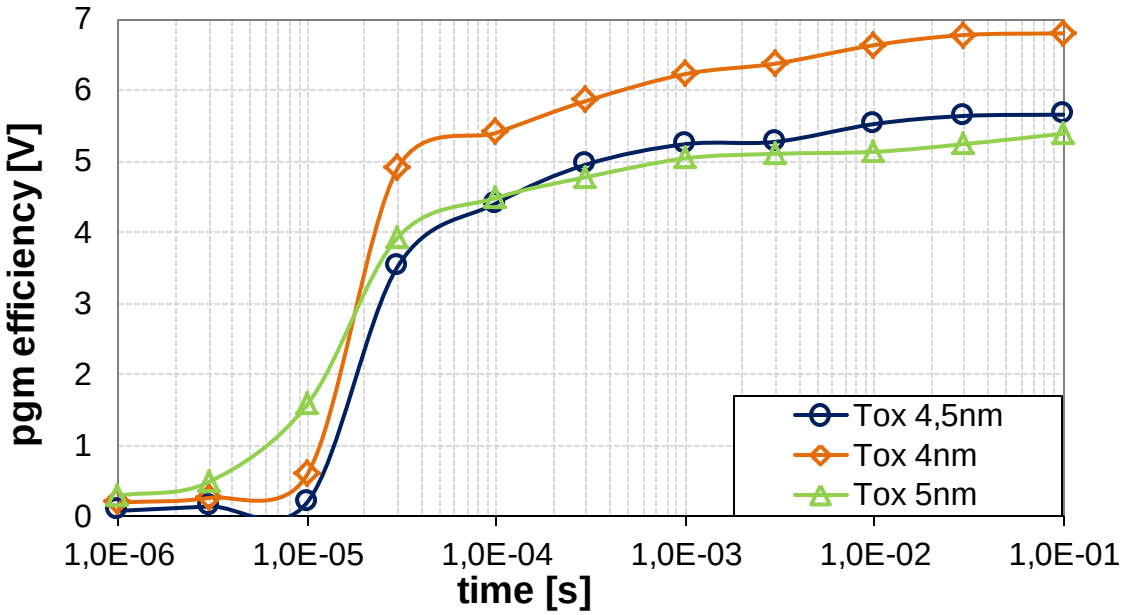


Figura 5.2.1: curve di efficienza di programmazione di celle SONOS floating body con area attiva tonda.

Le curve di programmazione misurate evidenziano il fatto che la cella risulta programmata dopo tempi di applicazione dell'impulso anche molto corti (10us).

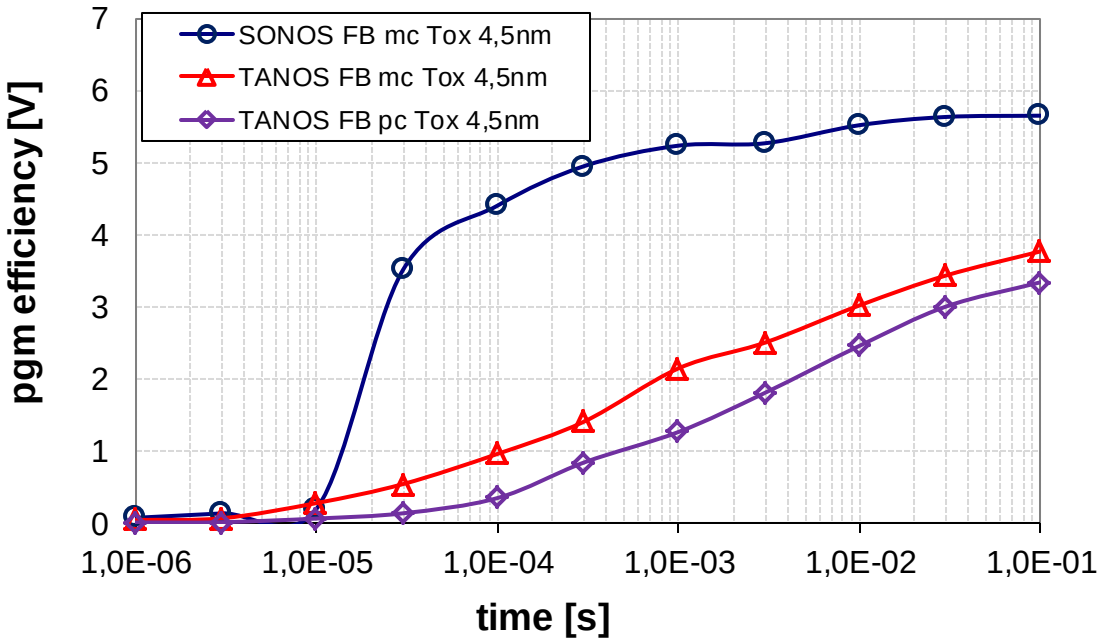


Figura 5.2.2: confronto di curve di efficienza di programmazione tra celle SONOS floating body e celle TANOS floating body (tensione di comando 16V).

In figura 5.2.2 è mostrato un confronto tra le efficienze di programmazione delle tecnologie viste finora a parità di condizioni operative. Come evidenziato nel precedente capitolo, si nota la maggior efficacia in programmazione delle celle TANOS con substrato in silicio monocristallino rispetto al policristallo; inoltre, gli elevati valori di efficienza a parità di durata dell'impulso applicato testimoniano come la forma del gate SONOS con area attiva tonda sia molto più incisiva negli effetti che regolano l'iniezione di carica dal substrato al nitruro. A parità di condizioni di lavoro, le finestre in programmazione delle celle SONOS floating body sono molto maggiori rispetto al caso TANOS.

5.3 Curve di cancellazione

Così come nel caso delle celle TANOS floating body, non avendo modo di contattare il substrato, anche per le SONOS floating body si sfrutta la corrente inversa che fluisce attraverso le giunzioni laterali in modo che la carica raggiunga il substrato polarizzandolo per una corretta cancellazione: per farlo, le giunzioni laterali devono

essere polarizzate inversamente. L'operazione di cancellazione è stata condotta applicando la tensione di comando simmetricamente a BL e SL, polarizzando a massa tutte le WL della stringa e ponendo i selettori DSL e SSL a 8V (per i motivi esaminati nel capitolo precedente).

Per quanto visto all'inizio del capitolo, l'operazione di cancellazione dovrebbe risultare molto più efficace rispetto a ciò che avviene per le classiche SONOS planari. Ammettendo una corretta polarizzazione del substrato, la differenza di potenziale ai capi dell'ossido di bottom genera un campo elettrico tale da estrarre le cariche dallo strato di trapping: poiché la forma del nitruro è arrotondata così come quella dell'ossido, il campo elettrico sarà molto più efficace. In effetti la figura 5.3.1 mostra quanto era atteso. Come si può osservare, per gli spessori di ossido più sottili si raggiungono elevati valori di efficienza anche per brevi (~100us÷1ms) tempi di applicazione della tensione di comando. E' utile notare che un'altra dipendenza è rispettata, ovvero quella che lega lo spessore dell'ossido di bottom all'efficienza di cancellazione, che aumenta con l'assottigliarsi del bottom oxide.

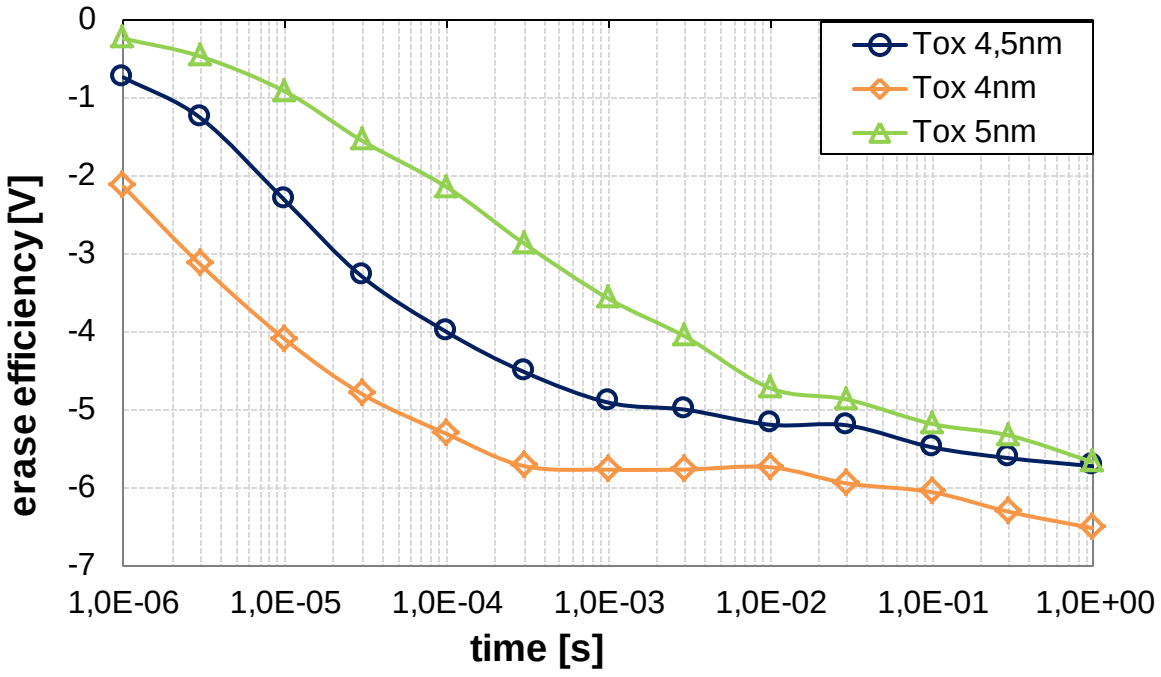


Figura 5.3.1: curve di efficienza di cancellazione di celle SONOS floating body con area attiva tonda.

In figura 5.3.2 è presentato il confronto a parità di condizioni operative con le tecnologie TANOS floating body analizzate nel capitolo precedente. Anche in questo caso si evidenzia come la modifica della forma dell'area attiva e degli strati del gate giochi un ruolo fondamentale nel miglioramento della finestra in cancellazione delle SONOS floating body.

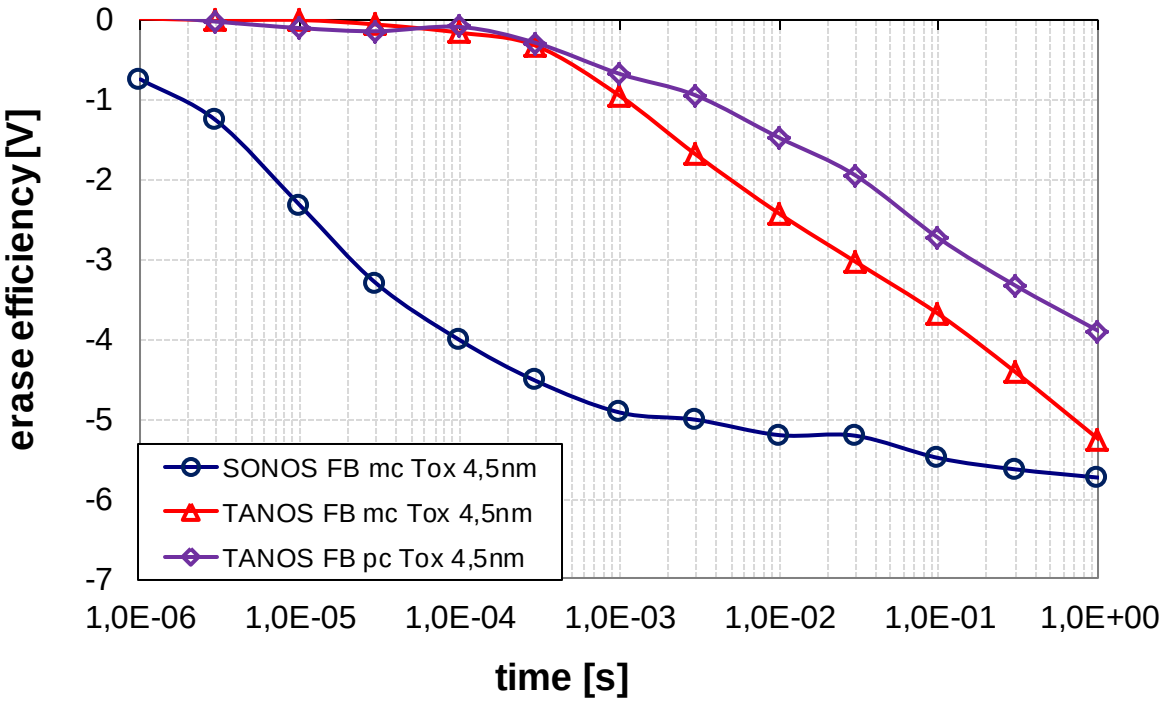
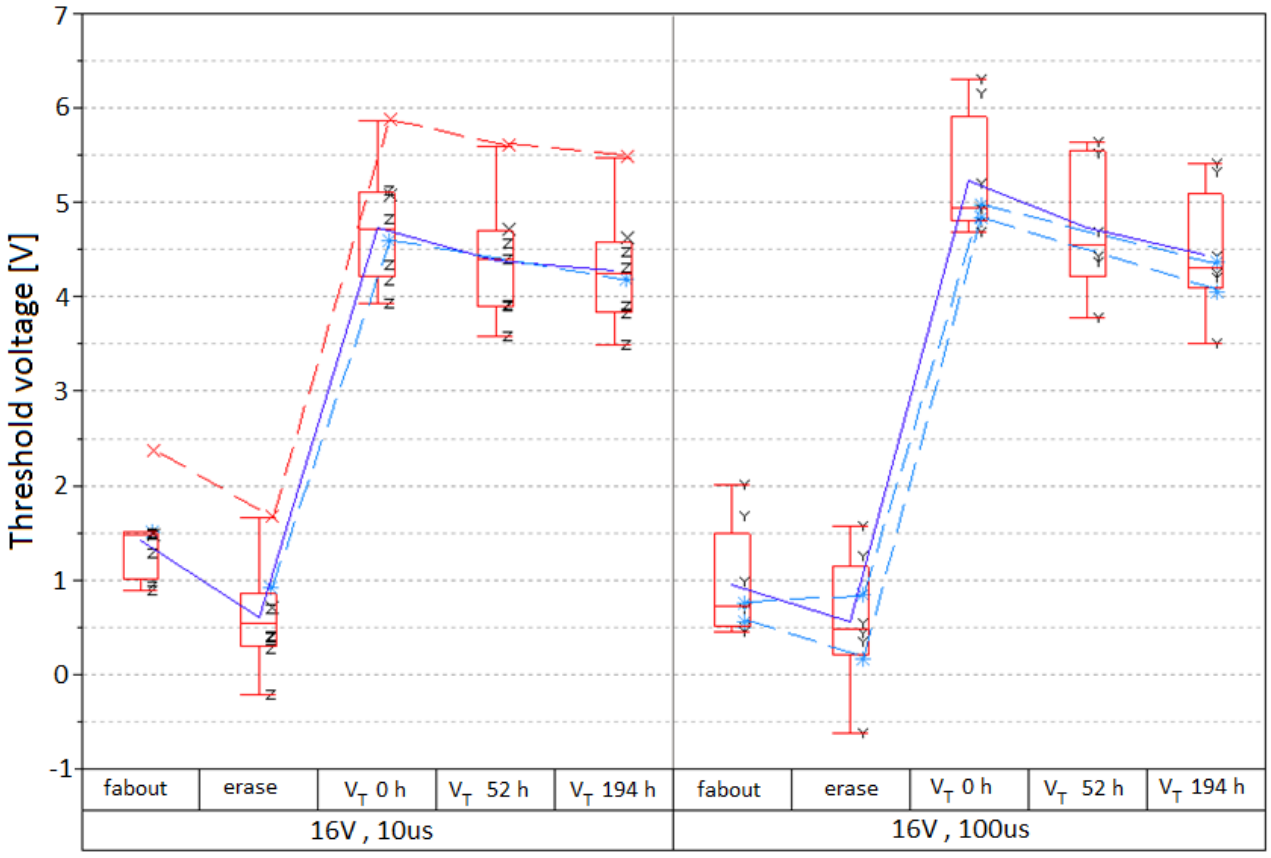


Figura 5.3.2: confronto di curve di efficienza di cancellazione tra celle SONOS floating body e celle TANOS floating body (tensione di comando 16V).

5.4 Ritenzione

Le prestazioni in ritenzione sono state valutate sperimentalmente effettuando una prima cancellazione in modo da svincolare la dipendenza delle operazioni dalle tensioni di soglia vergini delle celle, successivamente sono stati applicati segnali di comando per iniettare cariche nel nitruro in modo da poter estrapolare la variazione rispetto alla soglia intorno ad un valore di soglia programmata di 3,5V. Per un gruppo di celle sono stati applicati impulsi di programmazione da 16V di durata 10us, per un secondo gruppo di celle gli impulsi hanno la stessa ampiezza ma durata di 100us. Il wafer è stato successivamente inserito in un forno (bake) a 60° col fine di stimolare

l'emissione di carica dal nitruro per effetto termoionico e valutare le perdite con letture delle transcaratteristiche dopo periodi di tempo di 52 ore e 194 ore. I risultati delle misure sono stati riassunti e schematizzati in figura 5.4.1.



(a) (b)

Figura 5.4.1: andamento della tensione di soglia in ritenzione per celle SONOS floating body con area attiva tonda nelle due differenti condizioni di programmazione.

La perdita di carica dopo 52 ore dalla programmazione della cella comporta un abbassamento della tensione soglia di circa 0,41V: confrontando questo valore con il valore di una cella SONOS planare standard (dell'ordine di 0,15V) si può concludere che la cella a body flottante con area attiva tonda offre peggiori prestazioni dal punto di vista della ritenzione. Questo andamento è confermato anche dalle misure effettuate dopo 194 ore: valutando il degrado della pendenza con cui la tensione di soglia delle celle programmate con tempo 10us, essa è stimabile in -190mV/dec; un valore decisamente maggiore dei -120mV/dec con cui si stima la perdita delle classiche

SONOS planari. Si può dunque affermare che le prestazioni in ritenzione di queste strutture SONOS sono peggiorate rispetto alle architetture SONOS standard, proprio per la presenza dell'area attiva tonda, elemento che per controparte migliora l'efficienza di cancellazione.

5.5 Conclusioni

In questo capitolo abbiamo descritto le strutture SONOS floating body con area attiva tonda. Abbiamo descritto la particolare forma della cella e abbiamo visto come questo vada ad influire sulle prestazioni del dispositivo in fase di programmazione e di cancellazione. In particolare, è stata superata la problematica legata all'*erase saturation* delle classiche celle SONOS planari. La cella, per come è fatta, risulta più sensibile alla durata degli impulsi di comando che vengono applicati, mostrando un'elevata efficienza sia in programmazione che in cancellazione. Il prezzo da pagare per queste caratteristiche risiede nella minore capacità di ritenzione: questa caratteristica risulta degradata rispetto a quanto mostravano le classiche SONOS planari. In sostanza, nella misura di questi dispositivi SONOS, si osserva un trade off tra ritenzione ed efficienza di cancellazione: aumentare le prestazioni in un senso causa un notevole degrado nell'altro e viceversa.

Conclusioni

Oggetto del presente lavoro di tesi è stata la caratterizzazione sperimentale di celle di memoria per array 3D, evoluzione delle tradizionali memorie flash planari.
Nella parte introduttiva sono stati presentati i concetti di scaling e riduzione delle dimensioni dei transistors, di immagazzinamento dei dati in nodi di storage discreti per mezzo dell'idea del charge trapping, di sviluppo di memorie stacked e tridimensionali come miglioramento delle classiche tecnologie floating gate.
La prima parte dell'attività sperimentale ha caratterizzato le prestazioni delle memorie a substrato contattabile in cui il gate è stato modificato realizzando le strutture TANOS, alcune delle quali già analizzate in precedenza. In particolare sono stati passati in rassegna i risultati ottenuti nelle misure delle curve di programmazione e di cancellazione, delle ciclature e delle ritenzioni.
La seconda valutazione sperimentale ha riguardato le stesse caratteristiche su strutture differenti, memorie TANOS a substrato flottante. E' stato messo in evidenza come in questi dispositivi sia possibile raggiungere buone efficienze di cancellazione anche senza contattare direttamente il substrato, approfondendo il concetto tramite misure volte a verificare la simmetria dell'operazione di cancellazione sulla stringa SOI. Dal confronto tra le due tecnologie è emerso un sostanziale allineamento dei dati e delle prestazioni delle architetture analizzate, a testimonianza di come l'impossibilità di contattare il substrato non costituisca un limite nella realizzazione di dispositivi di memoria efficienti e performanti. I vantaggi della morfologia delle memorie TANOS a substrato flottante risiedono nella possibilità di sviluppare le stringhe su più piani e conseguentemente realizzare array di memorie 3D che possono aumentare la capacità di memorizzazione dei dati dei dispositivi elettronici.
Infine, nell'ultima parte, sono stati presentati i risultati delle misure effettuate su celle di memoria in tecnologia SONOS a substrato flottante ed area attiva tonda. Le

evidenze sperimentali hanno messo in luce come, ad un aumento dell'efficienza nelle operazioni di programmazione e cancellazione, corrisponda un peggioramento delle caratteristiche in ritenzione, imponendo quindi un trade off tra queste due caratteristiche della struttura.

Possiamo quindi affermare che le prestazioni delle celle TANOS e SONOS a substrato flottante sono comparabili con le tecnologie planari a substrato contattabile, con il vantaggio di poter realizzare array 3D per memorie stacked; tuttavia le evidenze sperimentali confermano performance inferiori rispetto alle classiche strutture floating gate.

Bibliografia

[1] R. Bez, E. Camerlenghi, A. Modelli e A. Visconti, *Introduction to Flash Memory*, Proceedings of the IEEE, vol. 91, no. 4, april 2003.

[2] W. D. Brown e J. E. Brewer, *Nonvolatile Semiconductor Memory Technology*, IEEE Press, 1998.

[3] S.K. Tewksbury et J.E. Brewer, Nonvolatile memory technologies with emphasis on Flash, 2008.

[4] Y. Taur T. H. Ning, *Fundamentals of modern VLSI technology*, pag. 164, 1998.

[5] T. Watanabe, *NAND Flash Memory Technology*, Toshiba Corporation, IEEE 2006.

[6] R. Entner, A. Gehring, H. Kosina, T. Grasser e S. Selberherr, *Impact of Multi-Trap Assisted Tunneling on Gate Leakage of CMOS Memory Devices*, NSTI-Nanotech 2005.

[7] K.J. Kuhn, *CMOS scaling beyond 32nm: challenges and opportunities,* IEEE Conferences, July 2009.

[8] J. Robertson e M.J. Powell, *Gap states in silicon nitride*, Appl. Phys. Lett. 44.

[9] Van den bosch, G.; Arreghini, A.; Breuil, L.; Cacciato, A.; Schram, T.; Suhane, A.; Zahid, M.B.; Jurczak, M.; Van Houdt, J.; *Understanding the impact of metal gate on TANOS performance and retention*, 2010

[10] William O'Leary, IBM *Advances Chip Technology With Breakthrough For Making Faster, More Efficient Semiconductors*, 1998

[11] E. K. Lai et al., IEDM 2006

[12] H. Tanaka et al., 2007 VLSI Tech.

[13] J. Kim et al., 2009 VLSI Tech.

[14] W. Kim et al., 2009 VLSI Tech.

[15] C. M. Compagnoni, A. Mauri, S. M. Amoroso, A. Maconi e A. S. Spinelli, *Physical Modeling for Programming of TANOS Memories in the Fowler-Nordheim Regime*, IEEE transactions on electron devices, Vol. 56, N. 9, 2009.

[16] C-H. Lee, C. Kang, J. Sim, J-S. Lee, J. kim, Y. Shin, K-T. Park, S. Jeon, J. Sel, Y, Jeong, B. Choi, V. Kim, W. Jung, C-I. Hyun, J. Choi e K. Kim, *Charge Trapping Memory Cell of TANOS (Si-Oxide-SiN-Al_2O_3-TaN) Structure Compatible to Conventional NAND Flash Memory*, IEEE 2006

[17] S.M. Sze, Kwok K. NG, *Phisycs of Semiconductor Devices*, pag. 314-316, Wiley, 3rd ed. 2006