POLITECNICO
MILANO 1863

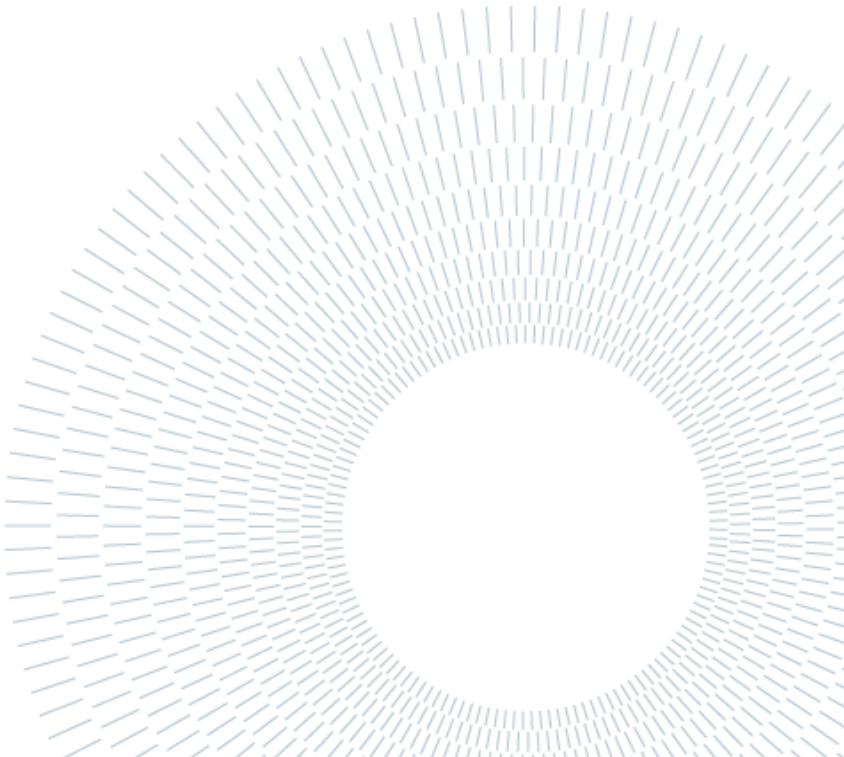
SCUOLA DI INGEGNERIA INDUSTRIALE
E DELL'INFORMAZIONE

Modello del DCP nelle memorie 3D NAND Flash

TESI DI LAUREA MAGISTRALE IN
ELECTRONICS ENGINEERING
INGEGNERIA ELETTRONICA

Autore: **Vito Francesco Moncelli**

Matricola: 940672
Relatore: Prof. Alessandro Sottocornola Spinelli
Anno Accademico: 2021-22



Sommario

Sommario

La crescente volontà di aumentare la densità di dati immagazzinabili, in dispositivi di dimensioni sempre più ridotte e più economici da fabbricare, ha condotto la ricerca verso la miniaturizzazione delle memorie NAND. Per anni, l'architettura planare della tecnologia 2D NAND ha subito la riduzione della dimensione caratteristica *F*. A questo è corrisposto l'abbattimento dei costi di produzione e l'aumento della capacità delle memorie, ma ha individuato anche alcuni limiti intrinsecamente insuperabili, dovuti all'effetto della natura discreta della carica in sistemi con *F* minore di 10nm. Non potendo contare più sullo scaling delle celle di memoria bidimensionali, è stata impiegata la terza dimensione con il conseguente avvento delle memorie 3D NAND Flash. Tramite una geometria innovativa, in cui i *control gates* (CG) modulano la concentrazione di carica in inversione circondando la regione di canale, è stato possibile assistere a un aumento della densità di dati già con valori di F abbastanza grandi da evitare i problemi di miniaturizzazione. Le convenzionali operazioni di programmazione, verifica, scrittura e cancellazione hanno richiesto un maggiore studio per via delle problematiche introdotte dalla struttura. In particolare, in questo lavoro di tesi, sarà analizzato il *down-coupling phenomenon* (DCP) avente luogo in seguito alla fase di verifica che procede la programmazione delle celle. Considerata la breve finestra temporale tra queste due fasi, il sistema non riesce a tornare all'equilibrio e la sua differenza di potenziale con le wordlines porta a disturbi di programmazione. Saranno investigati i meccanismi fisici che definiscono le dinamiche che portano il sistema all'equilibrio e si produrrà un modello circuitale col fine di stimare le dipendenze di questo fenomeno da alcuni parametri di progettazione.

Parole chiave: 2D NAND Flash, 3D NAND Flash, F, DCP, equilibrio, modello

Abstract

The growing need of increasing the density of storable datas in smaller devices, cheaper to produce, has led to the miniaturisation of the NAND memories. For many years, the feature size *F* of the planar architecture of the 2D NAND technology has been reduced. This has decreased the production cost, but it has brought attention to some unavoidable limits, due to the discrete nature of charge, especially in systems with *F* smaller than 10nm. Since a further scaling was unachievable, many resources where employed for the design of a new structure, able to use the third dimension, the 3D NAND Flash technology. Thanks to an innovative geometry, in which the control gates (CG) act on the inversion charge in the channel that they surround, it was possible to obtain an improvement in storage density without an excessive shrinking of *F*. The traditional operations, *program-verify-read-erase,* have required an additional care because of the problems caused by the new structure. Among them, in this work of thesis the *down-coupling phenomenon* (DCP) will be analyzed. It takes place after the verify operation the preceeds the program phase of the cell. Because of the brief time window between the two, the system can't reach an equilibrium state, so the potential drop between the channel region and the wordlines causes program disturbs. The goal of this study will be to detect the main mechanisms that define the dynamics that bring the system to equilibrium. Moreover a circuital model will be proposed in order to estimate the dependences on some design parameters.

Key words: 2D NAND Flash, 3D NAND Flash, F, DCP, equilibrium, model

Capitolo 1

Introduzione alle memorie NAND Flash

Negli scorsi anni la tecnologia NAND flash ha attirato l'attenzione del mercato per via dei vantaggi che offre. La resistenza allo stress meccanico e le capacità di cui possiede, spiegate nel seguito, l'hanno resa un'ottima alternativa agli hard disk drives. In questo capitolo sarà introdotta la tipologia 2D NAND, insieme alle modalità in cui opera. In seguito, partendo dai limiti legati alla miniaturizzazione, si offrirà una panoramica sulle motivazioni che hanno spinto alla creazione della tecnologia 3D NAND.

1.1 Avanzamento tecnologico

Negli ultimi anni, con l'aumento della quantità di dati immagazzinabili nelle memorie a stato solido, si è assistito alla crescita del mercato occupato da questa tecnologia. Infatti, questo fattore, legato alla ridotta quantità di silicio impiegata nella progettazione, ha influito sull'abbattimento dei costi di produzione. Il progresso scientifico ha fatto in modo che la tecnologia Flash diventasse un'alternativa affidabile, garantendo una capacità di contenimento di dati, maggiore di un $Gbit/mm^2$. Questo è stato possibile grazie alla costante miniaturizzazione delle strutture, il cui trend viene descritto in figura 1.1 dalla legge di Moore, secondo cui avviene una riduzione della dimensione caratteristica *F* (feature size) di un fattore $\sqrt{2}$ ogni due anni. Con i vantaggi ottenuti tramite la riduzione della dimensione delle memorie, sono sorti anche problemi dovuti alla natura discreta della carica e alle interferenze elettrostatiche. Per queste ragioni, le memorie 3D NAND hanno rappresentato una buona alternativa, permettendo un aumento del numero di dati senza un'eccessiva riduzione della dimensione *F*.

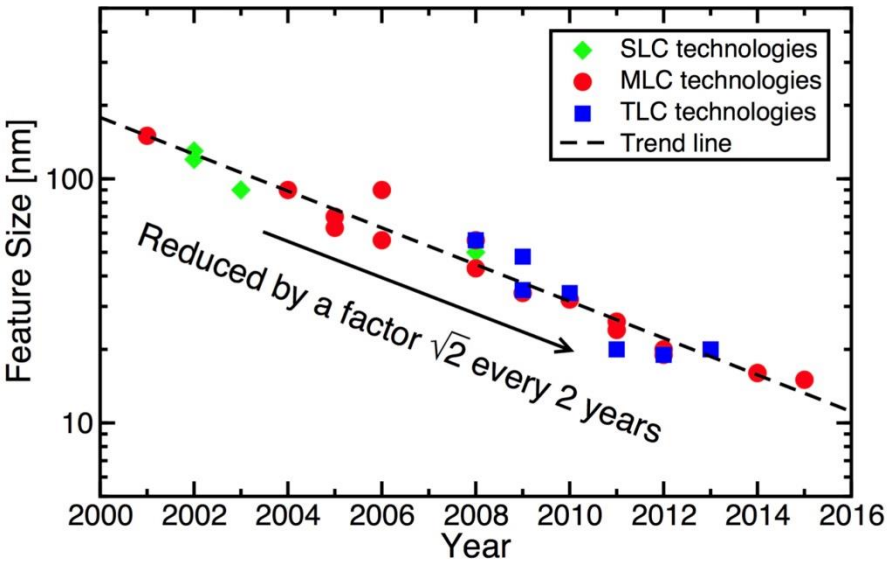


Figura 1.1 Andamento di *F* per le memorie 2D NAND dal 2001 al 2015

1.2 Memorie 2D NAND Flash

1.2.1 Principi di funzionamento

A partire dalla figura 1.2, in cui si riconoscono gli elementi fondamentali della struttura MOS, si nota un'aggiunta tipica di questa tecnologia. Infatti, oltre alla presenza di drain e source, con un drogaggio di tipo n, posizionati in un substrato di silicio di tipo p, è presente una sezione chiamata floating gate (FG). Questa parte della struttura, a differenza delle altre regioni, non ha una tensione fissata da un contatto e quindi rimane flottante. Il gate invece viene denominato control gate (CG) ed è separato da uno strato di ossido, il blocking oxide, dal floating gate. Per programmare la cella è necessario che una certa quantità di elettroni passi dal substrato al FG attraverso lo strato di ossido, tunnel oxide. Inoltre, è fondamentale che questa carica non vari nel tempo, prima dell'operazione successiva.

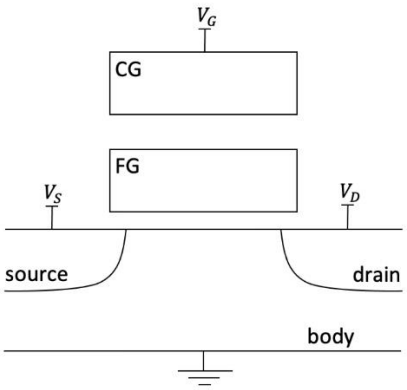
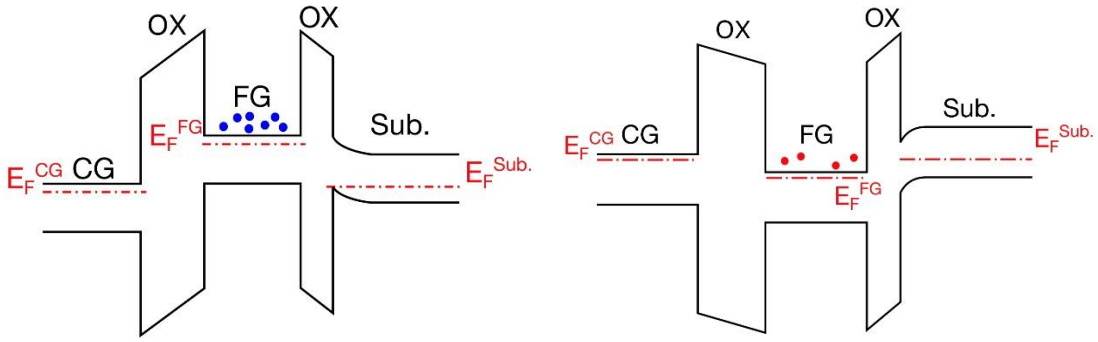


Figura 1.2 rappresentazione bidimensionale della cella di memoria



(a) (b)

Figura 1.3 diagramma a bande della memoria programmata (a) e cancellata (b)

In figura 1.3a viene raffigurato il diagramma a bande di una cella *programmata*, con elettroni nel floating gate, mentre in figura 1.3b si mostra quello di una cella *cancellata*, in presenza di carica positiva. La presenza di una carica negativa aumenta il potenziale elettrostatico, ostacolando quindi l'inversione di carica nel substrato, di conseguenza vi è un'incidenza sulla tensione di soglia (V_{th}) poiché è necessaria imporre una tensione maggiore tramite il control gate. Questo fenomeno viene impiegato per la distinzione dei due stati (erased e programmed). Tipicamente, con carica positiva e tensione di soglia bassa, si indica lo stato cancellato, a cui viene associato un livello logico pari a 1, nell'altro caso, con tensione di soglia maggiore, si indica il livello 0. A questa modifica di V_{th} corrisponde una modulazione della corrente, in particolare avviene uno shift rigido della caratteristica I_d-V_{CG} pari alla variazione della tensione di soglia, descritta dalla relazione 1.1, in cui q è la carica elementare,

C_{pp} è la capacità tra CG e FG e n_{fg} è la densità di carica intrappolata. Da queste considerazioni si deduce che, per conoscere lo stato logico della cella, è sufficiente applicare una tensione V_{read} e misurare la corrente I_d del drain.

$$\Delta V_{th} = -\frac{q \cdot n_{fg}}{C_{pp}} \tag{1.1}$$

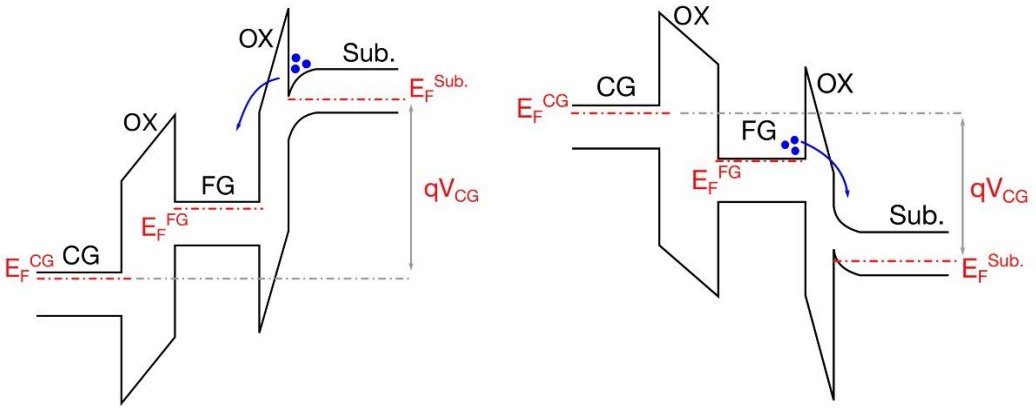


Figura 1.4 diagramma a bande durante l'operazione di programmazione (a) e nella fase di cancellazione (b)

1.2.2 Programmazione e cancellazione di una cella

Da quanto detto, risulta fondamentale controllare la quantità di carica nel floating gate per modificare lo stato logico della cella. A tal fine si può ricorrere all'effetto tunnel tramite il quale è possibile immettere elettroni nell'elettrodo. In presenza di una tensione molto alta V_{CG} del control gate e a patto che il tunneling oxide sia sottile, è possibile descrivere questo trasporto di carica tramite la relazione 1.2, che lega il flusso di elettroni (J_{FN}) al campo elettrico in regime Fowler-Nordheim (FN), per cui è richiesta una forma triangolare della barriera di potenziale vista dall'elettrone presente nel substrato.

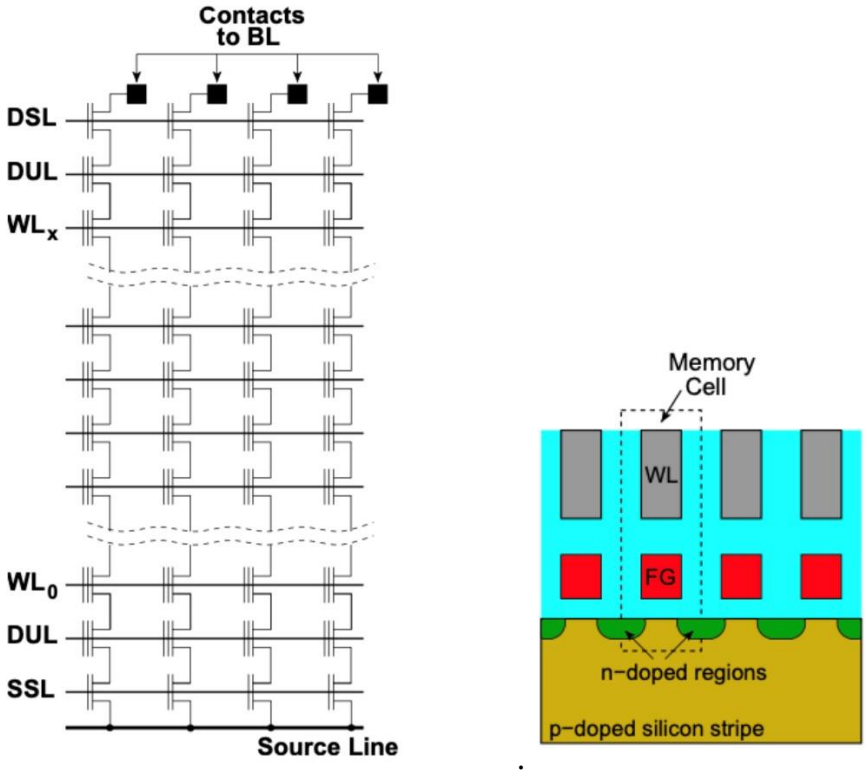
$$J_{FN} = A_{FN} F_{ox}^2 e^{\frac{-B_{FN}}{F_{ox}}} \tag{1.2}$$

Il valore assunto da A_{FN} e B_{FN} viene descritto da (1.3) e (1.4), in cui φ_B indica l'altezza della barriera di potenziale, m_{ox} la massa efficace di tunneling, mentre m_t e m_l indicano le masse efficaci trasversali e longitudinali.

$$A_{FN} = \frac{q^3(2m_t+4\sqrt{m_t\, m_l})}{16\pi^2 hq\varphi_B m_{ox}} \tag{1.3}$$

$$B_{FN} = \frac{4\sqrt{2\, m_{oxl}}}{3hq}(q\varphi_B)^3 \tag{1.4}$$

A partire da questo fenomeno, è possibile effettuare sia la programmazione della cella, sia la cancellazione, dissipando la minore quantità di energia possibile, in quanto non scorre una corrente tra drain e source. Si rende esplicito che per la fuoriuscita di elettroni va applicata una tensione V_{CG} negativa.



(a) (b)

Figura 1.5 (a) schema circuitale di un array di memoria 2D NAND Flash controllato in tensione tramite wordlines (WL). BL = bit-line, DSL = drain-select line, DUL = dummy line SSL = source-select line. (b) sezione planare dell'array.

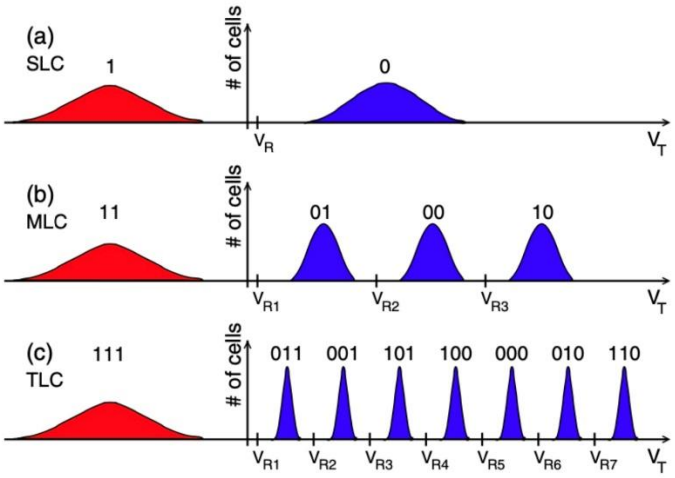


Figura 1.6 distribuzione della tensione di soglia per una cella a singolo livello (a), per una a multilivello (b) e una a triplo livello (c)

1.2.3 Struttura dell'array

Per immagazzinare più di un dato si pongono in serie più celle in modo tale da formare una stringa. Come si nota dalla figura 1.5 (a), ogni serie di celle è collegata da un lato ad una bitline specifica, mentre dall'altro tutte le stringhe sono connesse ad un'unica sourceline. Per abilitare la connessione con questi due contatti sono necessarie la drain select line e la source select line, entrambe precedute da una dummy cell che viene posta per evitare l'interferenza di campi elettrici indesiderati sulle celle interne. Per questa ragione, essa non viene usata e non è conteggiata per la stima della densità di dati della memoria. Perpendicolarmente si susseguono le wordlines che controllano una cella per ogni stringa e tra esse si distingue la dummy wordline, specifica per le omonime celle. Il parametro *F* citato in precedenza corrisponde alla dimensione longitudinale di una wordline. In figura 1.5 (b) si nota che il source e il drain di due celle adiacenti corrispondono e che per assicurare l'isolamento elettrostatico tra le celle è necessaria l'introduzione di un ossido.

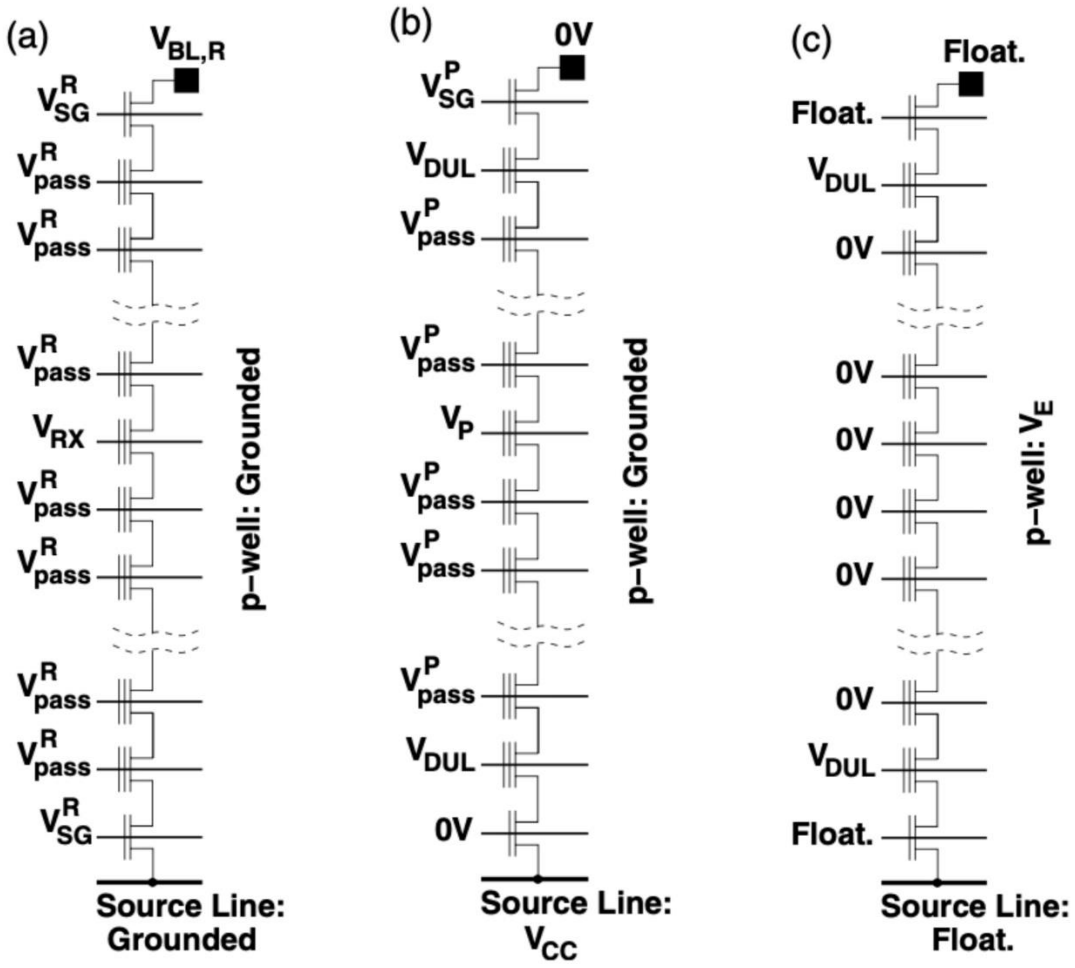


Figura 1.7 schema di polarizzazione per ogni operazione sull'array: (a) lettura, (b) programmazione, (c) cancellazione.

1.3 Operazioni sull'array

Partendo dalla relazione lineare tra variazione di tensione di soglia e carica intrappolata, è possibile definire più di un livello logico per ogni cella, uno per ogni tensione di soglia. Infatti, come osservato in figura 1.6, è possibile ottenere anche otto combinazioni, per una cella a triplo livello (TLC), considerando che il numero di stati è pari a 2^{BCP}, dove BCP (bits per cell) è la quantità di bit per ogni cella. Per BCP = 2, la cella è detta multilvel (MLC). In figura 1.7 viene mostrato lo schema di polarizzazione nei tre regimi descritti nel seguito.

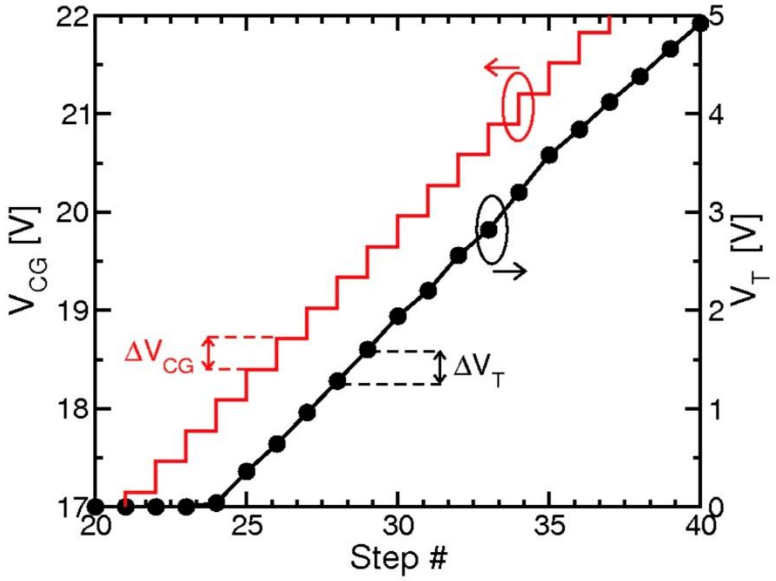


Figura 1.8 processo ISPP. Graduale aumento di tensione V_{CG} (in rosso) e conseguente varia di tensione di soglia (in nero)

Il primo processo analizzato è l'operazione di lettura (*read*) che consiste nell'individuazione della tensione di soglia della cella selezionata. Questo avviene tramite un circuito esterno connesso alla stringa che permette di confrontare il valore di V_{th} con una tensione di riferimento V_{RX}. A tal fine, si polarizza il control gate con la tensione V_{RX} e si pone la tensione V_{PASS} delle altre celle a un valore maggiore della massima tensione di soglia prevista dal sistema, in modo tale che non limitino la conduzione della cella in analisi. Un discorso simile si applica alla DSL e alla SLL. La sourceline rimane fissa a 0V, mentre la bitline ha una tensione positiva per far scorrere una corrente, che viene misurata e da cui si estrapola il valore di soglia della cella. Per le celle TLC e MLC, poiché ci sono più valori di V_{th} è necessaria la ripetizione di questo processo per più volte per poter distinguere quale stato logico è trattenuto. Questo fattore, insieme al lento ritorno all'equilibrio dovuto alle resistenze e alle capacità parassite delle wordlines e della bitline, rendono lento il processo di lettura. Per quanto riguarda la *programmazione*, come mostrato in figura 1.7 (b), è necessario creare un forte campo elettrico per favorire il processo di tunneling. Esso si ottiene ponendo la tensione della wordline della cella da programmare a un valore V_P di circa 20V, mentre le altre celle sono in fase di conduzione. La tensione di bitline, invece, è a 0V e fissa la tensione del bulk della stringa, in quanto DSL è alto, in modo tale che la carica possa fluire dal contatto al gate. Al contrario, SSL essendo a 0V, disaccoppia la sourceline dal substrato. Per evitare di creare campi elettrici troppo

intensi nel dielettrico, la tensione applicata alle dummy cells è più bassa rispetto a quella delle altre celle, ma garantisce comunque la conduzione. La necessità di calibrazione di V_{th} nelle MLC e nelle TLC ha richiesto la creazione di un metodo di programmazione con cui controllare gradualmente l'immissione di carica. Questo è stato ottenuto implementando la tecnica ISPP (*incremental step pulse programming*), che prevede il graduale aumento di tensione come mostrato in figura 1.8. Una tecnica simile viene utilizzata per l'operazione di cancellazione del dato nelle celle (*erase*), e prende il nome di ISPE (*incremental step pulse erase*). In questo caso è la tensione del substrato ad essere modificata e portata a tensioni negative. Questo tipo di operazione avviene contemporaneamente per tutte le celle, ponendo le wordlines a massa. Per evitare un eccessivo stress dovuto al campo elettrico, tutti gli altri contatti vengono lasciati flottanti, come visibile in figura 1.7 (c). Sia al termine di questo processo, sia in seguito alla fase di programmazione, ha luogo una fase di verifica della tensione di soglia delle celle (*verify operation*).

1.4 Problemi di miniaturizzazione

Una proprietà importante di cui deve godere una memoria è la capacità di mantenere un'informazione indipendentemente dal tempo per cui va conservata. A questo si aggiunge l'immunità alle operazioni a cui vengono sottoposte le altre celle, che non la riguardano direttamente. Con gli anni, aumentando il GBSD (*gross bit storage capacity*), si è dovuta ridurre la dimenisone caratteristica *F*. A questo scaling sono corrisposti errori durante le operazioni di memoria che hanno richiesto l'introduzione di codici di correzione degli errori, ECCs (*error correction codes*). L'errore ha luogo quando avviene una variazione imprevista della tensione di soglia desiderata. Ci sono tre tipi di errore. Il primo,l' *errore di scrittura (write errors)*, si presenta quando in fase di programmazione viene immessa una quantità di elettroni errata nella cella, che va a modificare il valore di V_{th}. La causa è da ricercare nelle fluttuazioni statistiche che caratterizzano l'effetto tunnel, originando il *program noise*, che definisce la variazione casuale di carica introdotta durante l'ISPP. Durante l'esecuzione di altre operazioni sulla stringa, possono verificarsi variazione di V_{th}, a causa dell'introduzione di carica indesiderata nel FG, come può avvenire durante la fase di lettura, per via delle alte tensioni applicate alle wordlines. Questo aspetto, così come le interferenze elettrostatiche tra celle vicine, rappresenta la tipologia di errori nominata *disturb errors*. Anche in assenza di operazioni sul dispositivo, possono sorgere effetti indesiderati. Ad esempio, il TAT (*Trap-assisted tunneling*) può favorire

uno scambio di carica, con la conseguente immissione di elettroni nel FG, tramite una trappola nel tunnel oxide, anche in assenza di polarizzazione della struttura. Questo difetto può causare anche altri effetti, come l'RTN (*random telegraph noise*) e il *charge detrapping*. Questa categoria prende il nome di *data retention errors*.

1.4.1 Program noise

Precedentemente nella trattazione, è stata quantificata la relazione tra variazione ΔV_T e carica immagazzinata, legate dalla capacità tra FG e CG. Alla riduzione della dimensione delle memorie è corrisposta una proporzionale diminuzione della capacità C_{pp}, che ha inficiato sempre di più le prestazioni dei dispositivi a causa dell'inversa proporzionalità con ΔV_T. Sono stati raggiunti valori tali da dover considerare l'effetto della granularità della carica, a partire dall'operazione di programmazione per la definizione della soglia, che quindi diventa una quantità statistica, legata alle fluttuazioni della carica. Per quantificare l'effetto, usando una distribuzione poissoniana per l'introduzione di carica, si ottiene la relazione 1.5.

$$\sigma_{\Delta V_T} \cong \sqrt{\frac{q<\Delta V_T>}{C_{pp}}} \tag{1.5}$$

1.4.2 Trap-assisted tunneling

L'assenza di difetti nei dielettrici prodotti durante il processo di fabbricazione è essenziale, in quanto, con la ripetitiva applicazione di campi elettrici molto intensi, a causa dello stress sulla struttura, possono crearsi delle trappole nello strato di ossido che possono portare a un trasferimento di carica indesiderato. Questo effetto viene denominato SILC (*Stress-induced leakage current*) e indica lo scambio di carica, tramite effetto tunnel, favorito da un difetto nell'ossido che modifica localmente la barriera di potenziale, come rappresentato in figura 1.9, e causa uno shift della soglia.

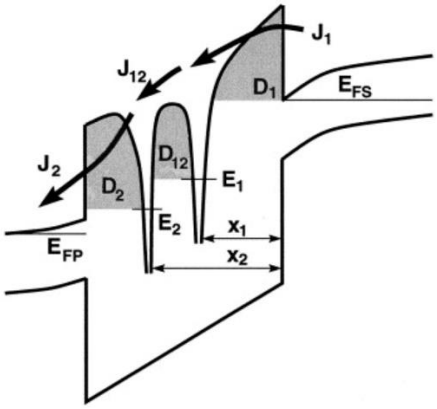


Figura 1.9 flusso di elettroni dovuto al trap-assisted tunneling

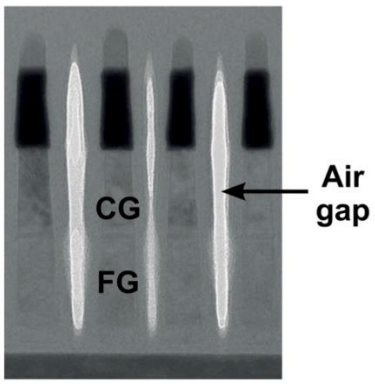


Figura 1.10 rappresentazione di air gaps

1.4.3 Interferenze tra celle

La riduzione del passo tra le celle ha portato a un maggior accoppiamento elettrostatico. Questo influenza le prestazioni di celle adiacenti, poichè in occasione della programmazione di una specifica cella, si assiste all'aumento di soglia anche delle celle limitrofe. Una soluzione a questo inconveniente è l'introduzione di aria tra celle contigue (air gap in figura 1.10), per abbassare la costante dielettrica e migliorare l'isolamento.

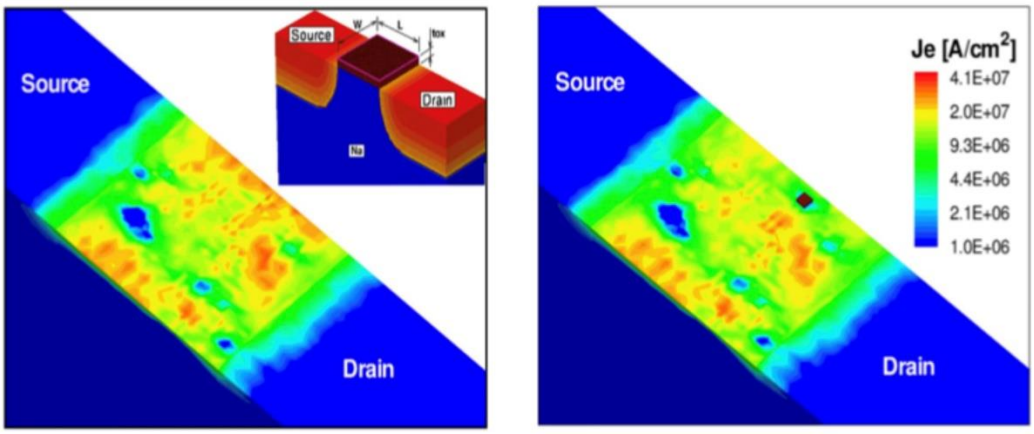


Figura 1.11 (a) densità di corrente di un MOSFET, simulata con drogaggio atomistico, (b) confronto della densità di corrente nel caso di un elettrone intrappolato

1.4.4 Random Telegraph Noise

I difetti all'interfaccia ossido-substrato possono causare *retention errors*, tramite il random telegraph noise (RTN). Questo effetto ha mostrato un'incidenza maggiore con lo scaling della tecnologia, in quanto è diventato sempre più rilevante valutare l'impatto delle singole cariche sulle prestazioni del dispositivo. L'RTN ha luogo a causa dell'effetto di un elettrone intrappolato all'interfaccia, poiché la sua presenza modifica l'elettrostatica locale e modula la corrente, dato che essa scorre in prossimità dell'ossido. Questa variazione di corrente rappresenta una variazione della tensione di soglia e diventa sempre più rilevante con la miniaturizzazione dei dispositivi, dal momento che la natura atomistica del drogaggio facilita la creazione di flussi di carica più ostacolabili dai difetti interfacciali.

1.4.5 Charge detrapping

Anche l'occupazione delle trappole nel tunnel oxide è da includere nell'analisi, infatti, può variare durante i cicli di programmazione e cancellazione e, sebbene non influisca sul SILC, può impattare sulla tensione di soglia. Se, ad esempio, una trappola ha una carica durante la fase di programmazione/cancellazione e invece diventa neutra nella fase di ritenzione, si assiste a una diminuzione della soglia.

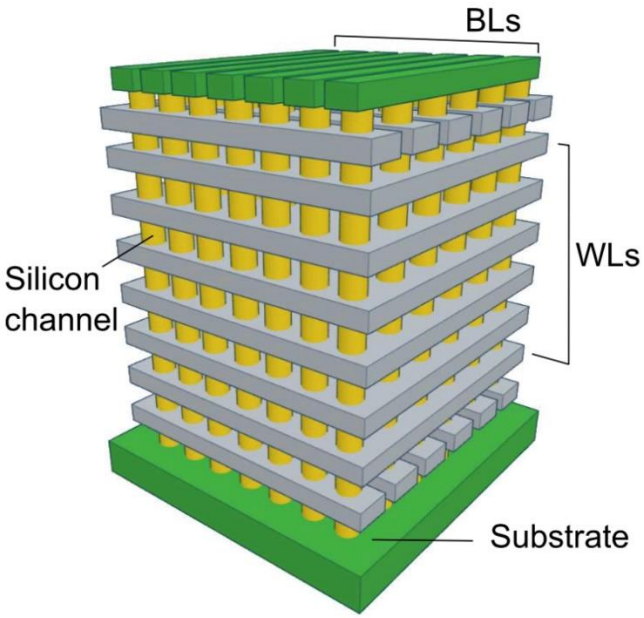


Figura 1.12 struttura di una memoria a canale verticale

1.5 Memorie 3D NAND Flash

Le problematiche legate alle memorie planari e la crescente richiesta di aumento di GBSD hanno portato ad un profondo interesse per le memorie 3D NAND. Esse garantiscono un aumento del numero di dati, senza la minimizzazione eccessiva del parametro F per la tecnologia precedente. L'idea di fondo dietro la creazione di questa generazione di memorie è l'impiego della terza dimensione, quella verticale, per aumentare la quantità di dati a parità di superficie. La progettazione risulta però complessa e ha portato a due tipi di sistemi: *Vertical channel memories*, su cui sarà posta l'attenzione, e *Vertical gate memories*.

1.5.1 Struttura a canale verticale

Come nelle memorie planari, viene conservata l'idea di bitline, sourceline e wordline, ma esse vengono integrate in maniera diversa. In figura 1.12, ogni cella è data dall'intersezione di una wordline, che non è più una linea, bensì un piano, con la struttura cilindrica (*pillar*) in cui si crea il canale. Il pillar corrisponde al bulk della struttura bidimensionale e ad ognuno di essi corrisponde una stringa, collegata ad una sourceline comune. La bitline invece è specifica per

ogni riga di stringhe. In figura 1.12 si riconoscono delle wordline dirette perpendicolarmente rispetto alla direzione delle bitline, e servono per abilitare una singola stringa.

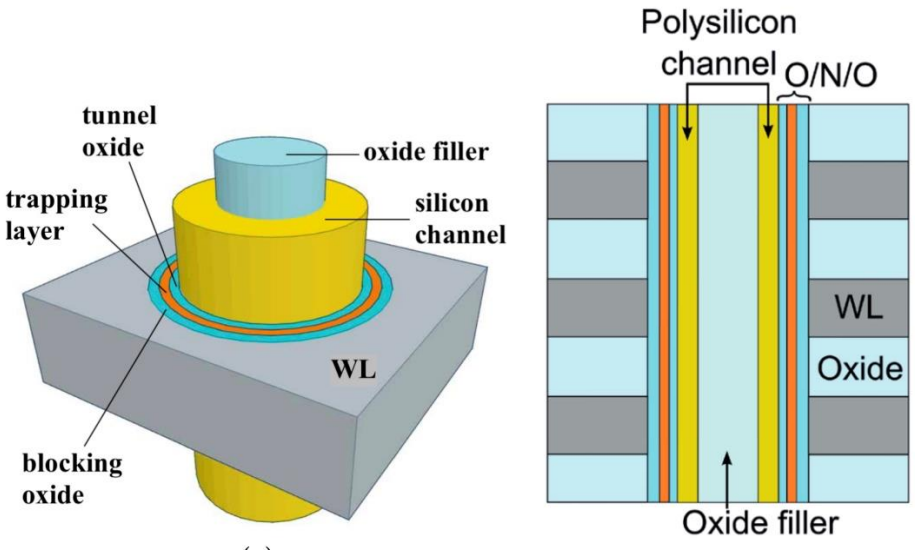


Figura 1.13 tecnologia GAA (a) cella singola (b) sezione verticale della memoria tridimensionale

In figura 1.13 (a) si nota che un miglior controllo elettrostatico del canale nel pillar è garantito per una cella GAA (*gate all around*). L'elettrodo flottante delle memorie 2D viene rappresentato da un materiale dielettrico con elevata densità di difetti, al fine di ridurre fenomeni di correnti di leakage. Per ridurre ulteriormente questo problema, rappresentato dal trasporto di carica tramite il tunnel oxide e il blocking oxide, si è ricorsi all'ONO, composto da $SiO_2 - Si_3N_4 - SiO_2$, che offre una buona barriera isolante per contrastare il buon accoppiamento elettrostatico tra wordline e canale. Per evitare correnti tra le wordline, in polisilicio fortemente drogato, è stata necessaria l'introduzione di un ossido di separazione. Il processo tecnologico alla base di questa struttura prevede il deposito dell'ONO all'interno del pillar, poi a causa del processo di integrazione non è possibile ottenere silicio monocristallino, quindi si crea il polisilicio. Questo materiale è caratterizzato dalla presenza di grani con piani di orientazione sfasati. Alle interfacce tra queste regioni, la mancanza di legami atomici agisce come trappola per gli elettroni. Nella parte centrale si trova il *Macaroni (oxide filler)* la cui introduzione serve a impedire il congiungimento dei grani al centro della struttura, che implicherebbe una concentrazione maggiore di trappole ed è utile anche per ridurre la regione di canale, per evitare che le wordline abbiano un effetto di modulazione bidimensionale sulle varie parti della cella. Il tipo di struttura presa in analisi ha il nome di BICS (*Bit-Cost Scalable*) memory.

1.6 Caratteristiche delle memorie tridimensionali

Come citato in precedenza, la qualità di queste memorie risiede nella capacità di offrire un più alto GBSD anche per dimensioni maggiori di F, riducendo quindi i problemi legati allo scaling. Di conseguenza si ha anche una riduzione legata al program noise, all'RTN e ai fenomeni di charge detrapping, dovuti agli effetti relativi alla granularità della carica che diventano dominanti nelle memorie 2D. A questo corrisponde un maggiore controllo sulla variabilità della tensione di soglia. Il problema principale risiede nell'elevata densità di difetti introdotti dall'uso del polisilicio, che hanno come effetto un aumento della resistenza di canale che riduce la corrente di lettura e introduce variabilità per la distinzione di V_{th}. Le interfacce tra i grani favoriscono inoltre la dipendenza dell'RTN dalla temperatura ed errori legati al charge detrapping.

Capitolo 2

Down-coupling phenomenon

In questo capitolo saranno forniti più dettagli sull'effetto delle operazioni svolte sulla stringa e particolare attenzione sarà posta sul fenomeno del downcoupling (DCP). Successivamente, tramite alcune simulazioni TCAD, sarà effettutato uno studio quantitativo delle dinamiche coinvolte a seguito di questa operazione

2.1 Operazione di programmazione

L'operazione di programmazione prevede due fasi, la verifica e la programmazione della cella tramite ISPP. In figura 2.1 viene mostrato l'andamento temporale dei contatti, da cui si nota un intervallo di tempo, tra *verify* e *program*, in cui tutte le tensioni sono poste a massa per favorire il ritorno all'equilibrio del sistema. Nonostante questo, la successione delle due operazioni può portare a problemi, soprattutto nelle stringhe non abilitate, che non dovrebbero modificare il loro stato anche in presenza dell'alta tensione di wordline. Per descrivere gli errori che sorgono, di seguito si analizza in dettaglio la fase di programmazione.

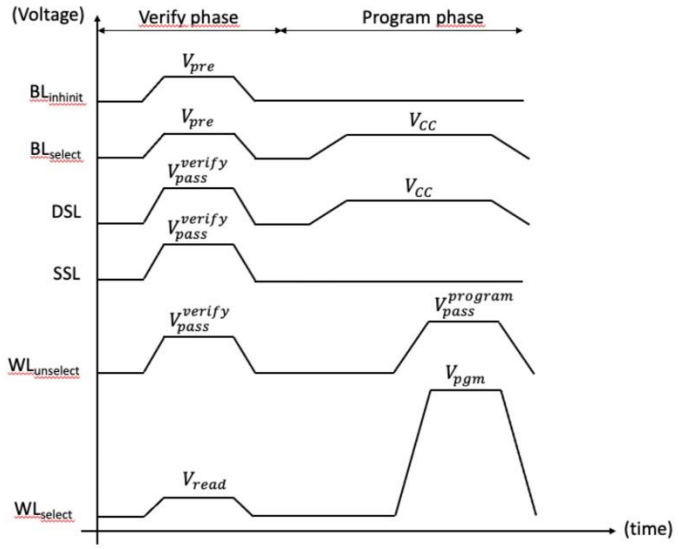


Figura 2.1 schema di polarizzazione durante la programmazione

2.2 Effetto di self-boosting

Per evitare la programmazione indesiderata di celle non selezionate, è necessario un'inibizione del processo di immissione di elettroni. Questo si può ottenere riducendo il campo elettrico tra canale e gate delle celle non selezionate. Per far verificare questa eventualità si fa affidamento alle capacità gate-canale e si tiene a massa la SSL. La DSL invece è posta a una tensione V_{cc} prima della programmazione, in modo tale che, applicando V_{cc} anche alla bitline, la regione di canale si trovi a una tensione $V_{cc} - V_T^{DSL}$. Così facendo, il transistor controllato dalla DSL risulta spento, la stringa flottante e il potenziale della stringa subisce una variazione riportata in 2.1, tramite un accoppiamento capacitivo con le WLs.

$$\Delta V_{CH} = \frac{C_{WL-CH}}{NC_{tot}}(N-1)V_{pass} + \frac{C_{WL-CH}}{NC_{tot}}V_{pgm} \qquad (2.1)$$

Sebbene la tecnologia 3D NAND favorisca il disaccoppiamento con bitline e sourceline, grazie all'assenza del contatto al centro della stringa, viene misurata comunque un'alta differenza di

potenziale in prossimità della wordline a tensione V_{pgm}. Infatti, dalla relazione 2.1, si nota che la tensione V_{pass} è quella che definisce maggiormente la tensione.

2.2.1 Effetto di local self-boosting

Da quanto detto nella sezione precedente, traspare che all'aumentare di V_{pass} migliora la tecnica di inibizione di programmazione, di contro non può essere usata una tensione eccessivamente alta, per evitare disturbi sulle celle non selezionate nella stringa da programmare. Di conseguenza, persiste il problema dell'eccessiva differenza di potenziale tra gate e canale delle celle della stessa wordline della cella da programmare.

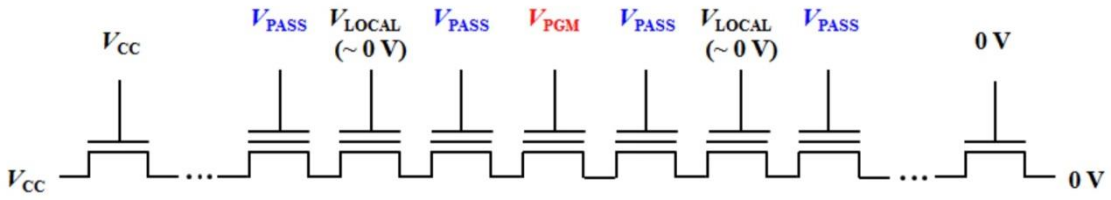


Figura 2.2 schema di polarizzazione per il local self-boosting

Una soluzione è il local self-boosting, per cui si portano a 0V le wordline di due celle intorno alla cella da programmare. Così facendo, si causa lo spegnimento dei transistor e il conseguente disaccoppiamento di un numero M di celle, racchiuse tra le wordline a 0V, dal resto della stringa. La nuova variazione di potenziale è descritta da 2.2 e lo schema di polarizzazione di questa operazione viene mostrato in figura 2.2, per M = 3.

$$\Delta V_{CH} = \frac{C_{WL-CH}}{NC_{tot}}(N-1)V_{pass} + \frac{C_{WL-CH}}{MC_{tot}}V_{pgm} \qquad (2.2)$$

2.3 Down-coupling phenomenon

Prima della programmazione di una cella, ha luogo un'operazione di verifica, che avviene analogamente alla lettura, alzando la tensione di bitline e ponendo le wordline a una tensione V_{pass}^{verify}. A causa del DCP, fenomeno descritto di seguito, la stringa può trovarsi a una tensione negativa, causando *disturb errors*. Questo fenomeno ha luogo in seguito alla rapida variazione delle wordline da V_{pass}^{verify} a 0V e consiste nell'abbassamento di potenziale della stringa in seguito allo spegnimento dei transistor delle celle non selezionate.

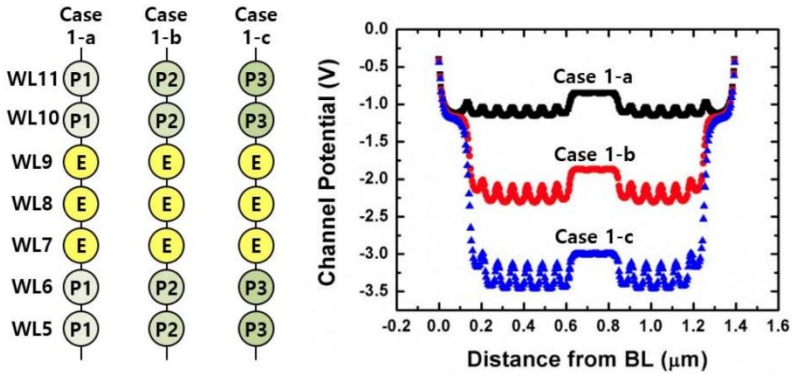


Figura 2.3 effetto del DCP per diverse tensioni di soglia delle celle non selezionate

Durante questo processo, la variazione di tensione del gate della cella opera una modulazione di carica nel canale solo fino alla tensione di soglia, poiché controlla la concentrazione di elettroni in banda di conduzione. Per tensioni minori, la dinamica della stringa è dipendente solo dalla naturale evoluzione del sistema verso l'equilibrio. Come per il fenomeno di self-boosting, anche in questo caso, la tensione a cui si trova la regione di canale è dipendente principalmente dalle celle non selezionate, poiché sono in numero maggiore, quindi, si può modellare questa variazione tramite la relazione 2.3.

$$\Delta V_{Down-coupling} \approx V_{th,celle} \tag{2.3}$$

In figura 2.3 si osserva la variazione di potenziale di una stringa da 16 celle in seguito al DCP, da cui si nota che questo effetto ha luogo indipendentemente da quale sia la tensione di soglia

delle celle programmate. Questo fenomeno è fortemente dipendente dalla scelta dei parametri della regione di canale, infatti, a un drogaggio elevato corrisponde una diminuzione di lacune, la cui presenza è richiesta per il raggiungimento dell'equilibrio. In più, come mostrato nel seguito di questo lavoro di tesi, anche le trappole giocano un ruolo fondamentale. Sebbene sia presente una fase di alcuni microsecondi in cui tutti i contatti vengono posti a zero, per favorire il ritorno all'equilibrio, le dinamiche in gioco sono più lunghe e possono causare *errori di disturbo* nella fase di programmazione. Il seguente lavoro si concentrerà sull'individuazione dei processi necessari per il raggiungimento dell'equilibrio. Nel seguito, saranno indicati i dettagli inerenti alle simulazioni TCAD per lo studio di questo fenomeno. Inizialmente, verrà presentato l'ambiente di lavoro, in seguito sarà posta l'attenzione sulla geometria della struttura e sui fenomeni fisici coinvolti nella simulazione.

2.4 Ambiente di simulazione

Per eseguire uno studio quantitativo del DCP, si è ricorso all'uso di un *software TCAD* chiamato *Sentaurus,* prodotto da *Synopsys*. Dapprima, la riproduzione virtuale della struttura geometrica è stata realizzata tramite *Sentaurus Structure Editor*. Successivamente, la definizione del tipo di simulazione, i parametri numerici e i modelli fisici sono stati implementati in *Sentaurus device, tool* destinato alla risoluzione delle equazioni indicate dall'utente. La selezione dei modelli ha incluso *Drift-Diffusion transport model, B2BT* (Band-to-band tunneling), *Avalanche generation model, SRH* (Schockley-Read-Hall) *generation-recombination process*. In aggiunta, la scelta di simulare l'impatto delle trappole sul DCP ha richiesto l'introduzione di modelli come il *TAT* (*trap-assisted tunneling*) e l'effetto *Poole-Frenkel*.

2.5 Geometria della stringa di memoria

La natura cilindrica della struttura presa in analisi consente, per ragioni di simmetria, lo studio di una sezione bidimensionale longitudinale per la comprensione dei fenomeni che hanno luogo. Inoltre, poiché la geometria della parte centrale è periodica, e dal momento che la connessione di questa regione ai contatti laterali avviene secondo le medesime modalità, è possibile concentrare l'analisi solo su metà della struttura.

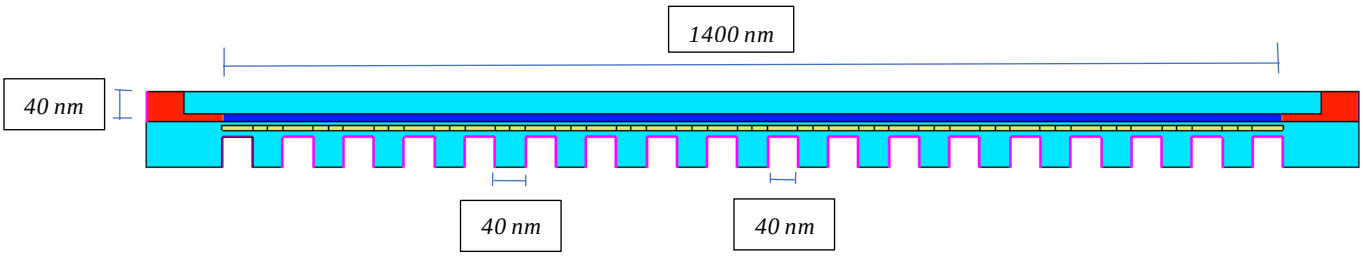


Figura 2.4 Sezione longitudinale della stringa con 16 celle di memoria

La stringa su cui è stata effettuata l'analisi, riportata in *figura 2.4*, dispone di *16* celle di memoria e *2 select transistors*, distanziati longitudinalmente da uno strato di ossido *(SiO2)* di *40nm*. Spostandosi lungo la direzione radiale, si riconosce l'*ONO layer*, caratterizzato da uno spessore di *8nm* sia per il *blocking oxide (SiO2)*, sia per la regione di *silicon nitride (Si3N4)*, mentre da uno spessore di *4nm* per il *tunneling oxide (SiO2)*. La regione di canale, in *polisilicio*, si estende longitudinalmente per *1400nm* e radialmente per 10nm, e agli estremi è collegata alle regioni di polisilicio di *sourceline (SL)* e di *bitline (BL)*. Esse sono state definite con un drogaggio di tipo n, con una concentrazione di atomi di Arsenico pari a $5 \times 10^{19} cm^{-3}$. Scelta analoga è stata presa per la regione centrale di polisilicio, selezionando però un drogaggio di $10^{15} cm^{-3}$. Infine, si precisa che il materiale dell'*oxide filler (Macaroni)*, presente nella parte più interna del cilindro, è *SiO2*.

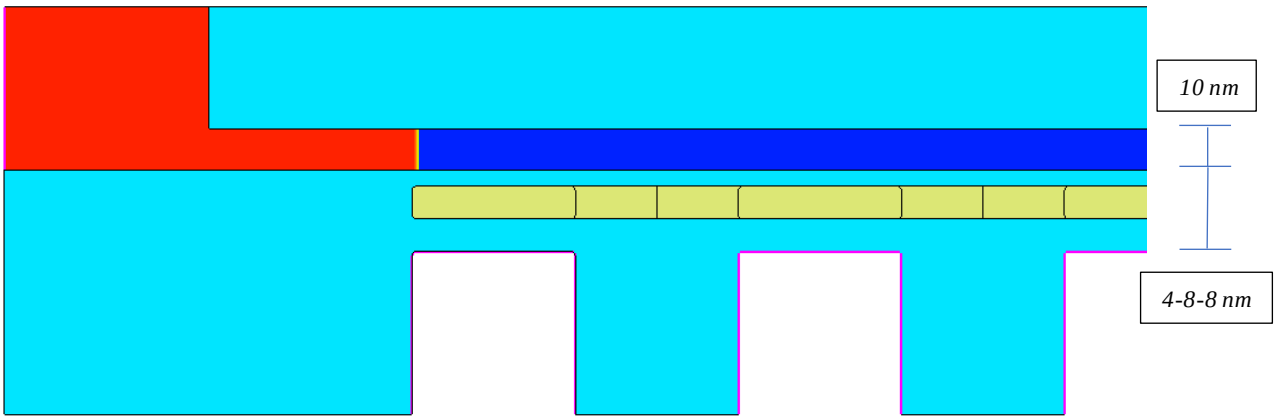


Figura 2.5 ingrandimento della sezione longitudinale della stringa

In *figura 2.5* si può identificare, in rosso, la regione del contatto di *bitline* e in blu la regione di canale, col colore celeste si indica invece la presenza di ossido *(SiO2)* e col giallo il *nitruro (Si3N4)*. Di seguito, in *tabella 2.6* vengono elencati i parametri geometrici. L_{tot} corrisponde alla lunghezza totale della struttura, R_{mac} rappresenta il raggio del filler cilindrico (*Macaroni*), L_{WL} è la misura della dimensione longitudinale del contatto della *wordline* e della separazione tra le celle, ed è pari anche alla lunghezza del contatto di *bitline*. R_{on} è l'estensione radiale del *blocking oxide* e del *nitruro*, R_s ed R_{ch} descrivono rispettivamente il raggio del *tunneling oxide* e della regione di canale.

R_{mac}	30nm
L_{WL}	40nm
R_{ch}	10nm
R_{on}	8nm
R_s	4nm
L_{tot}	1.6μm

Tabella 2.6 parametri geometrici

2.6 Modelli e parametri fisici della simulazione

2.6.1 Caratterizzazione delle trappole

In questa sezione vengono approfondite le motivazioni dietro la scelta dei modelli fisici integrati e si offre una panoramica dell'impatto che hanno i parametri numerici che caratterizzano questi fenomeni nelle simulazioni TCAD.

$$N_{a,t}(E) = N_{a,t}e^{(E-Ec)/k_bT} + N_{a,d}e^{(E-Ec)/k_bT} \quad (2.7)$$

$$N_{d,t}(E) = N_{d,t}e^{(E-Ev)/k_bT} + N_{d,d}e^{(E-Ev)/k_bT} \quad (2.8)$$

Una volta definiti la geometria e i materiali del sistema, è stato necessario quantificare la concentrazione e la distribuzione delle trappole in banda proibita, nelle regioni in polisilicio. L'occupazione degli stati è stata descritta secondo l'*equazione 2.7 e 2.8,* mentre i livelli energetici delle trappole sono stati modellati tramite due distribuzioni esponenziali, sia per le trappole con comportamento *accettore*, sia *donore*. Per chiarezza, si fa presente che col termine *accettore* si indica un tipo di trappola non carica quando vuota, e con carica equivalente ad un elettrone quando occupata. Con *donore*, in caso di occupazione, si indica invece una carica pari a una lacuna. Inoltre, si è ritenuto opportuno procedere modellando le grandezze in questione in modo spazialmente costante, in tutto il polisilicio.

$N_{a,t}$	$K_t \times \sum_i N_{0a,t} e^{(-\lvert Ec-E_i\rvert/E_{a,t})}\ cm^{-3}$
$N_{a,d}$	$K_t \times \sum_i N_{0a,d} e^{(-\lvert Ec-E_i\rvert/E_{a,d})}\ cm^{-3}$
$N_{d,t}$	$K_t \times \sum_i N_{0d,t} e^{(-\lvert E_i-Ev\rvert/E_{d,t})}\ cm^{-3}$
$N_{d,d}$	$K_t \times \sum_i N_{0d,d} e^{(-\lvert E_i-Ev\rvert/E_{d,d})}\ cm^{-3}$
K_t	*1*
eXA	$10^{-16}\ cm^2$
hXA	$5 \times 10^{-15}\ cm^2$
eXD	$5 \times 10^{-15}\ cm^2$
hXD	$10^{-16}\ cm^2$

Tabella 2.9 parametri di caratterizzazione delle trappole

Nella tabella 2.9, viene indicato il modo in cui è stata implementata la distribuzione degli stati e vengono indicate le *cross section* sia per gli elettroni (eXD, eXA), sia per le lacune (hXA, hXD), per entrambe le tipologie di trappola. Modellati gli stati in banda proibita, è stato calcolato il loro impatto sulla variazione della tensione di soglia *(V_{th})* del dispositivo, definita misurando la tensione applicata al gate delle celle programmate tale da far scorrere nella stringa una corrente I_{BL} di *10nA*, ponendo la *bitline* a una tensione di *1V*, come mostrato in figura 2.10. Sono state simulate varie casistiche, imponendo che tutte le celle della stringa fossero programmate, modificando dapprima il parametro K_t , che modifica la concentrazione totale di trappole, ed in seguito la quantità di carica *(nP)* presente nel *nitruro*.

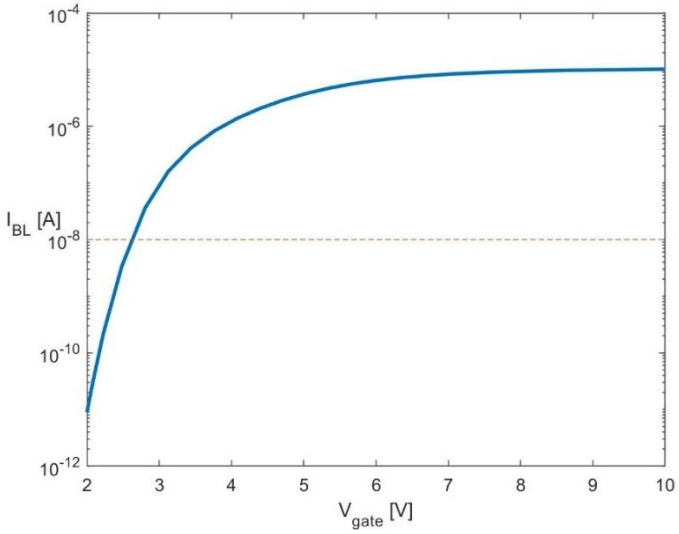


Figura 2.10 estrapolazione soglia per $nP = -10^{19}\ cm^{-3}, K_t = 1$

Estrapolato il primo valore, mantenendo invariata la concentrazione di trappole, è stata calcolata la tensione di soglia al variare della carica nel nitruro, aspettandosi una dipendenza lineare (Figura 2.11), come nel caso senza trappole.

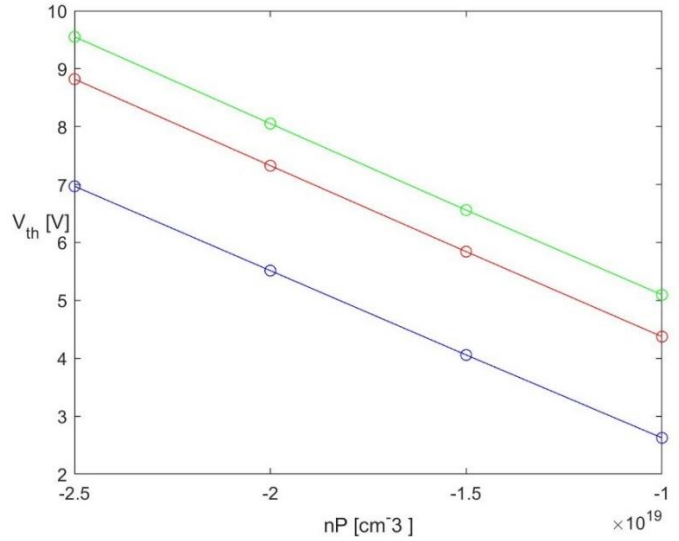


Figura 2.11 valori di tensione di soglia in funzione della carica di programmazione, per $K_t = 1$ *(blu), per* $K_t = 16$ *(rosso), per* $K_t = 25$ *(verde)*

Successivamente, si è valutato il valore V_{th} al variare della concentrazione di stati in banda proibita, tramite il parametro K_t. Quest'ultima operazione ha evidenziato una forte dipendenza di V_{th} da K_t (Figura 2.11), giustificata dall'ipotesi che, a parità di tensione applicata ai *gates* delle celle, gli elettroni intrappolati aumentano con K_t, riducendo la quantità di carica presente in banda di conduzione che può sostenere una corrente di *10nA*.

$nP\ [cm^{-3}]$	$V_{th}\ [V]$
-10^{19}	*2.626*
$-1.5\times\ 10^{19}$	*4.058*
-2×10^{19}	*5.514*
$-2.5\times\ 10^{19}$	*6.971*

Tabella 2.12 V_{th} in funzione di nP per K_t = 1

$nP\ [cm^{-3}]$	$V_{th}\ [V]$
-10^{19}	*4.374*
$-1.5\times\ 10^{19}$	*5.842*
-2×10^{19}	*7.326*
$-2.5\times\ 10^{19}$	*8.820*

Tabella 2.13 V_{th} in funzione di nP per K_t = 16

$nP\ [cm^{-3}]$	$V_{th}\ [V]$
-10^{19}	*5.096*
$-1.5\times\ 10^{19}$	*6.559*
-2×10^{19}	*8.052*
$-2.5\times\ 10^{19}$	*9.552*

Tabella 2.14 V_{th} in funzione di nP per K_t = 25

Osservato l'effetto dell'aumento della concentrazione di trappole e della variazione della carica nel nitruro, è stato necessario calibrare un ulteriore parametro di simulazione, il numero di *livelli di discretizzazione* della distribuzione esponenziale degli stati (N_L). A tal proposito, si fa presente che il simulatore usa, come impostazione predefinita, 13 livelli, mentre, nelle simulazioni condotte, questo parametro è stato modificato per comprendere se avesse un impatto significativo sulla tensione di soglia. I risultati forniti da *sdevice* hanno mostrato una dipendenza trascurabile dal parametro N_L per tutta la gamma di K_t e nP considerati. In particolare, nel caso di interesse, considerando $K_t = 1$ e una densità di carica nel nitruro pari a $-2\times\ 10^{19} cm^{-3}$, è stata registrata una variazione di V_{th} di circa 1% da $N_L = 25$ ($\mathrm{V_{th}} = 5.47$ V) a $N_L = 100$ ($V_{th} = 5.535\ V$).

2.6.2 Drift-Diffusion Model

Il *Drift-Diffusion model* descrive il trasporto dei portatori liberi di carica, assumendo che la corrente sia legata alla variazione spaziale del potenziale elettrostatico e al gradiente di concentrazione di lacune ed elettroni. Prendendo in esame l'espressione matematica più basilare, il parametro più significativo del modello *Drift-diffusion* è la *mobilità* (μ). Essa, pur essendo dipendente dal campo elettrico e dal drogaggio, è stata ritenuta costante in tutto il dispositivo, con un valore $\mu_n = 150\ cm^2/(V \cdot s)$ e $\mu_p = 120\ cm^2/(V \cdot s)$. Questi valori, inferiori rispetto al caso del silicio monocristallino, modellano la diminuzione di mobilità tra i grani del polisilicio, a causa della struttura policristallina. A causa della variabilità dell'orientazione e della dimensione dei grani del materiale, non è stata effettuata un'analisi approfondita sulla caratterizzazione di questo parametro. Si è optato per questa decisione, in primo luogo, al fine di concentrare l'analisi principalmente sulla dipendenza del DCP dalla presenza di trappole e, in secondo luogo, poiché si è ritenuto più centrale lo studio di altri meccanismi fisici dominanti. Di seguito vengono riportate le espressioni della densità di corrente di lacune ed elettroni (J_n , J_p), nel caso in cui è possibile legare la *diffusività* (D) alla *mobilità*, tramite la relazione di *Einstein* (*2.17*, *2.18*).

$$J_n = -nq\mu_n \nabla\varphi_n \qquad (2.15)$$

$$J_p = -pq\mu_p \nabla\varphi_p \qquad (2.16)$$

$$D_n = K_b T \mu_n \qquad (2.17)$$

$$D_p = K_b T \mu_p \qquad (2.18)$$

Con n e p, si indicano rispettivamente la densità volumetrica di elettroni e lacune, con q la carica elementare, con φ_n e φ_p i *quasi-fermi potentials*, con K_b la costante di Boltzmann e con T la temperatura.

2.6.3 Generation-Recombination process

In questa sezione vengono introdotti altri due maccanismi legati alla generazione e ricombinazione di carica: Band-to-band tunneling e Schockley-Read-Hall recombination. Il primo modella la generazione di lacune e elettroni in presenza di forte campo. Particolare attenzione va prestata al modello Schockley-Read-Hall: esso, infatti, stima la generazione e la ricombinazione di elettroni in banda di conduzione e di lacune in banda di valenza, legando questi meccanismi alla concentrazione, alla distribuzione e alle caratteristiche degli stati in banda proibita. In particolare, l'equazione 2.19, tramite i parametri τ_n e τ_p, rivela un'inversa proporzionalità alla densità dei difetti e alla sezione di cattura delle trappole e tiene conto della loro distribuzione tramite n1 e p1 (2.20, 2.21). In condizioni di equilibrio termodinamico, per la legge di azione di massa, il prodotto della densità di elettroni (n) e della densità di lacune (p) è pari al quadrato della concentrazione intrinseca dei portatori di carica (n_i^2), di conseguenza, il numeratore assume valore zero e non ci sono processi di generazione e ricombinazione.

$$R = \frac{np - n_i^2}{\tau_p(n+n1) + \tau_n(p+p1)} \tag{2.19}$$

$$n1 = n_i e^{\frac{E_{trap}}{K_b T}} \tag{2.20}$$

$$p1 = n_i e^{-\frac{E_{trap}}{K_b T}} \tag{2.21}$$

In 2.20 e 2.21, Etrap indica la distanza tra il livello del difetto e il livello intrinseco. In previsione di campi elettrici molto intensi, si è ritenuto necessario integrare i fenomeni di trap-assisted tunneling e di emissione Poole-frenkel. Questi fenomeni, infatti, non possono essere trascurati per campi elettrici superiori a $3 \times 10^5 {}^{V}/_{cm}$ e poiché si verificano prima degli effetti di B2BT e generazione a valanga, la loro introduzione nella simulazione è necessaria. Il simulatore, partendo dall'espressione 2.19, modella questi effetti riducendo i tempi di vita medi (τ_n e τ_p). In particolare, ricorrendo all' hurkx trap-assisted-tunneling model, si ottiene l'espressione 2.22, in cui γ_{TAT} è un parametro dipendente dal campo elettrico.

$$\tau = \frac{\tau_0}{(1+\gamma_{TAT})} \tag{2.22}$$

Nel caso dell'effetto Poole-frenkel, il modello prevede un aumento della probabilità di emissione (γ_{PF}) dai centri di carica intrappolata, dove avviene un abbassamento della barriera di potenziale a causa del campo elettrico esterno. Il software interviene direttamente sulla cross section (2.23), dove k è un fattore di scaling, modellando γ_{PF} (2.24) tramite un parametro α(2.25), dipendente dal campo elettrico (F).

$$\sigma = \sigma_0(1 + k\gamma_{PF}) \tag{2.23}$$

$$\gamma_{PF} = \frac{1}{\alpha^2}[1 + (\alpha - 1)e^{\alpha}] - \frac{1}{2} \tag{2.24}$$

$$\alpha \propto \sqrt{F} \tag{2.25}$$

2.7 Analisi dei risultati di simulazione

In questa sezione vengono riportati i risultati delle simulazioni effettuate e vengono evidenziati quali sono i meccanismi fisici dominanti per il raggiungimento dell'equilibrio del sistema. In aggiunta, si cerca di generalizzare e confermare le ipotesi proposte, anche per sistemi con maggiore numero di celle di memoria. Al termine della trattazione, al fine di ridurre gli effetti del DCP in fase di progettazione, si cerca di valutare e quantificare come alcune variazioni della geometria analizzata influenzino le prestazioni. Il risultato della prima simulazione, in cui sono stati abilitati tutti i modelli descritti con l'aggiunta di generazione tramite impatto ionizzante e B2BT, viene raffigurato in figura 2.26, dove viene riportato l'andamento energetico della banda di conduzione del centro della stringa ($L_{tot}/2$), in funzione del tempo. L'asse dei tempi parte da $10\mu s$, perché è stata impostata una variazione della tensione dei gate delle celle da 8V a 0V, con una funzione rampa a pendenza negativa di durata $10\mu s$.

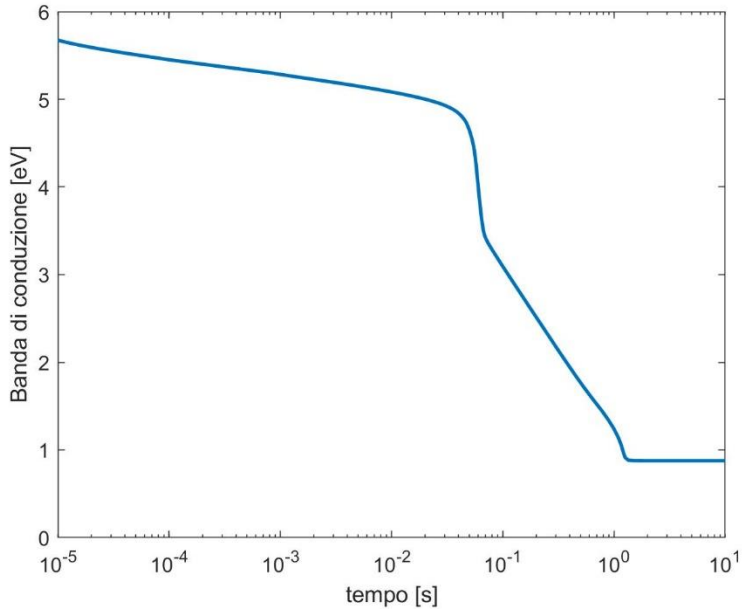


Figura 2.26 andamento della banda di conduzione al centro della struttura in funzione del tempo, con tutti i modelli fisici abilitati

Questi primi calcoli hanno dato un'indicazione sulla durata della dinamica del sistema, ma, per l'identificazione dei meccanismi dominanti, è stato necessario effettuare la medesima analisi

transitoria disabilitando alcuni modelli e valutando eventuali variazioni nel raggiungimento dell'equilibrio.

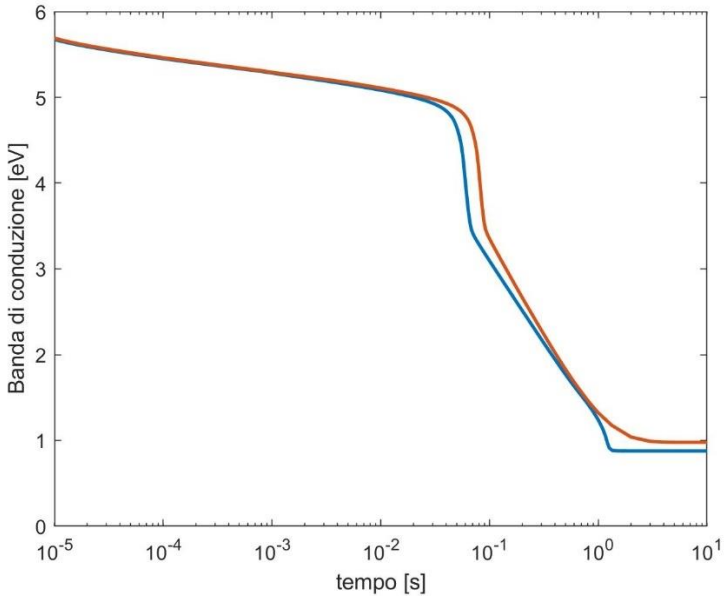


Figura 2.27 confronto dell'andamento della banda di conduzione al centro della struttura in funzione del tempo, con avalanche e B2BT abilitati (blu) e disabilitati (rosso)

Nella simulazione riportata in figura 2.27 sono stati disabilitati i modelli *avalanche* e *B2BT*. Le trascurabili variazioni della curva hanno suggerito una debole dipendenza da questi meccanismi, rafforzando l'idea che i meccanismi legati agli stati in banda proibita fossero determinanti. In particolare, ulteriori calcoli hanno dimostrato la centralità del trap-assisted-tunneling e dell'emissione Poole-Frenkel, fenomeni senza i quali è stata osservata una notevole dilatazione del tempo di raggiungimento del valore di equilibrio, come mostrato in figura 2.28.

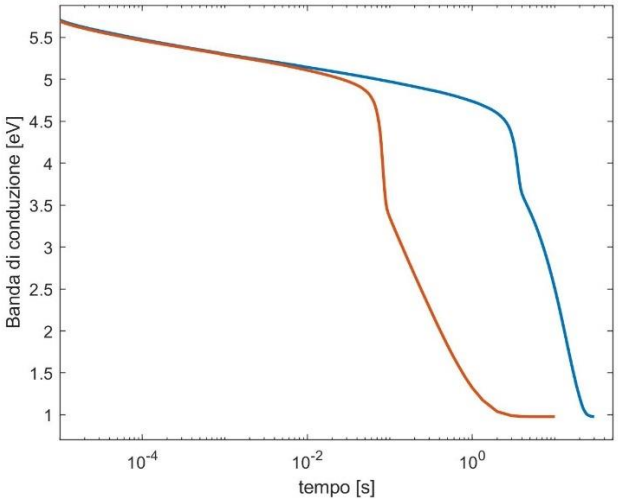


Figura 2.28 confronto dell'andamento della banda di conduzione al centro della struttura in funzione del tempo, con TAT e PF abilitati (rosso) e disabilitati (blu)

Fatte queste considerazioni, per spiegare i vari cambi di pendenza della curva, è stato utile valutare la concentrazione di carica intrappolata e libera in vari punti della stringa, nei punti significativi del grafico. In figura 2.29, viene mostrata la densità totale di elettroni e lacune nella regione di canale.

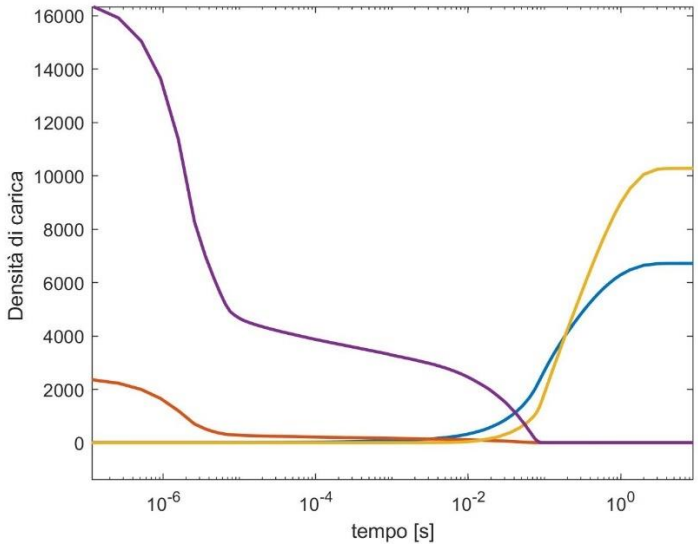


Figura 2.29 confronto delle densità di carica (μm^{-1}): elettroni intrappolati (viola), elettroni liberi (rosso), lacune intrappolate (giallo), lacune libere (blu).

Si nota una prima rapida variazione degli elettroni in banda di conduzione, da 0s a 10 μs, che corrisponde alla variazione delle wordlines da 8V a 0V. Dalla figura si comprende che l'abbassamento dei gate non è sufficiente a forzare l'uscita di tutta la carica in eccesso, infatti, la modulazione della carica della regione di canale è rilevante solo per valori maggiori rispetto al valore di soglia. Questo ha come effetto l'abbassamento di potenziale della stringa di circa -Vth. Per quanto riguarda la carica intrappolata, si assiste a un comportamento simile rispetto alla carica libera: in questa prima fase del transitorio, riescono a scaricarsi velocemente solo i difetti prossimi alla banda di conduzione, con costanti di tempo piccole. La carica di lacune è trascurabile quindi non influenza l'elettrostatica del sistema. Si ricorda che in questa struttura non ci sono contatti di *tipo p* che possono sostenere una importante immissione di lacune e che la regione dei contatti di *tipo n* raggiunge l'equilibrio in un tempo nell'ordine del *dielectric relaxation time* (picosecondi).

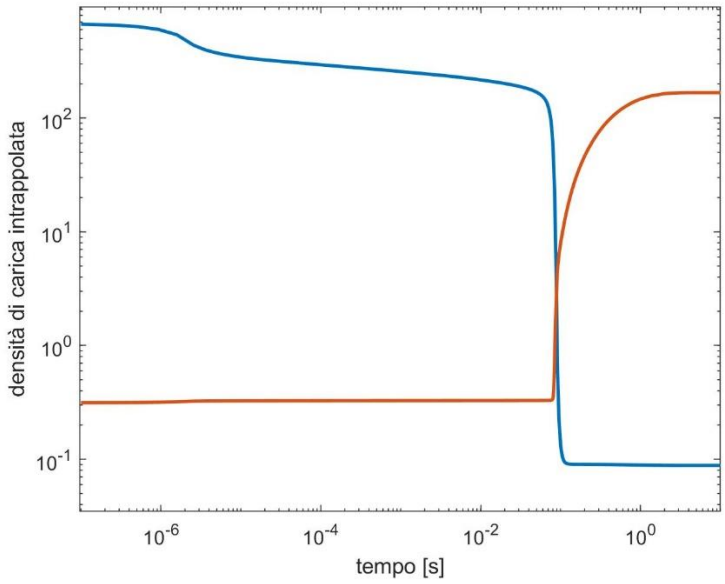


Figura 2.30 confronto delle densità di carica (μm^{-1}) nella regione centrale della struttura. elettroni intrappolati (blu), lacune intrappolate (rosso)

Mentre il lento di rilascio di carica da parte degli stati accettori provoca una leggera variazione dell'energia della banda di conduzione, come mostrato in figura 2.30, il forte campo elettrico, localizzato agli estremi della stringa, favorisce la generazione di lacune che tentano di

raggiungere il centro della struttura. Quando, in corrispondenza di una cella, la concentrazione di lacune diventa paragonabile a quella degli elettroni, avviene una rapida ricombinazione, con un conseguente abbassamento dell'energia della banda di conduzione e questo genera un forte campo elettrico che favorisce TAT ed emissione PF con la cella adiacente più interna, incrementando leggermente la generazione di lacune. In (2.31), si può notare l'abbassamento della banda di conduzione rispetto al centro.

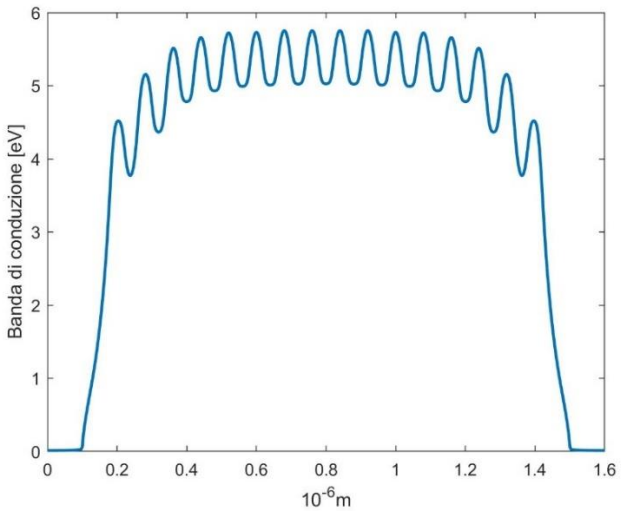


Figura 2.31 profilo longitudinale della banda di conduzione, all'istante t = 0.01 s

Dopo pochi istanti, solo la carica di lacune intrappolate sostiene il potenziale e si nota un abbassamento dell'energia della banda di conduzione costante per tutte le celle, come visibile in figura 2.32. Infatti, se le lacune tendessero a fermarsi nelle celle più esterne, si creerebbero un campo elettrico e un gradiente di concentrazione che assicurerebbero una rapida distribuzione di carica verso il centro. A questo punto, poiché il campo elettrico è debole, la produzione di lacune prosegue in maniera significativa solo ai bordi, fino al raggiungimento dell'equilibrio.

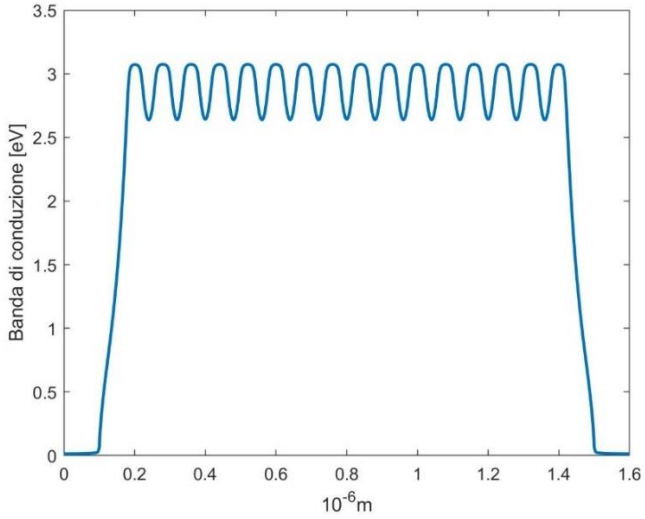


Figura 2.32 profilo longitudinale della banda di conduzione, all'istante $t = 0.2$ s

Vista la forte dipendenza dai fenomeni di generazione e ricombinazione, è stato valutato anche l'impatto delle variazioni di N_L sul transitorio.

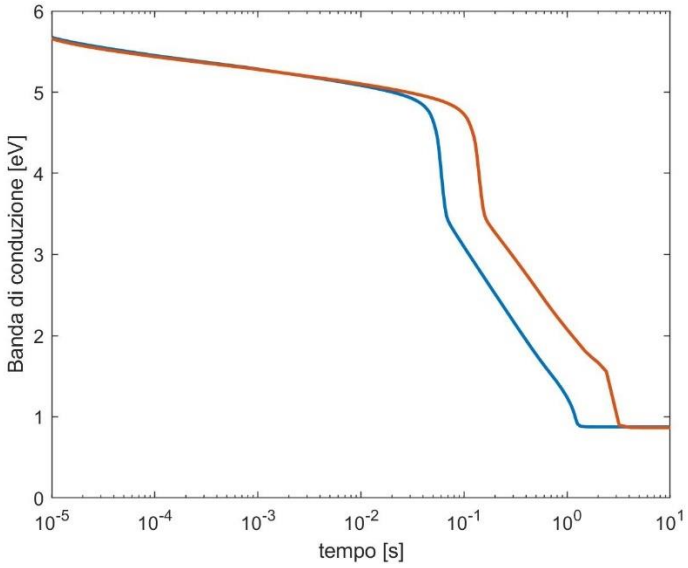


Figura 2.33 andamento della banda di conduzione al centro della struttura in funzione del tempo, con tutti i modelli fisici abilitati. $N_L = 50$ (rosso), $N_L = 75$ (blu)

L'immagine 2.33 mostra la medesima analisi, modificando il parametro N_L. Nonostante la concentrazione degli stati rimanga costante, all'aumentare di N_L corrisponde un maggior numero di livelli più prossimi a metà banda proibita, e poiché, questi ultimi danno il maggior

contributo nei processi di generazione, si nota un'evoluzione più veloce del sistema. Questo risultato implica una forte dipendenza dalla posizione delle trappole e quindi la difficoltà di prevedere le prestazioni di un dispositivo. Per completare la trattazione, sono stati calcolati gli effetti introdotti dall'aumento del numero di celle e dallo scaling della dimensione longitudinale dei gate e del distanziamento delle celle di memoria.

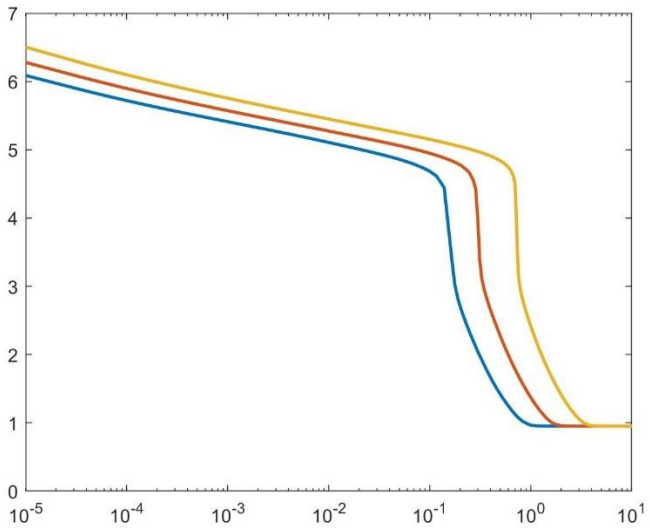


Figura 2.34 andamento della banda di conduzione al centro della struttura in funzione del tempo, per numero di celle (N) variabile: N = 64 (blu), N = 128 (rosso), N = 256 (giallo)

La figura 2.34 mostra un aumento pressochè lineare del tempo di equilibrio t_0 (10% del valore finale) con il numero di celle. La motivazione dietro questa dipendenza è legata alla generazione di lacune ai bordi della stringa. Infatti, in tempi molto minori rispetto a t_0, il potenziale nella stringa è circa costante e il campo elettrico tra le celle è circa zero. A questo corrisponde una generazione di lacune praticamente nulla nella regione di canale e tutta l'immissione di carica è affidata al TAT e all'emissione Poole-Frenkel ai bordi. Questi fenomeni, però, sono legati esclusivamente al campo elettrico agli estremi, che conserva un andamento uguale indipendentemente dal numero di celle e quindi vi è un'immissione di carica invariata, a fronte di una carica richiesta proporzionale al numero di celle. Quindi, ad esempio, raddoppiando il numero di celle si osserva circa un aumento di t_0 del doppio.

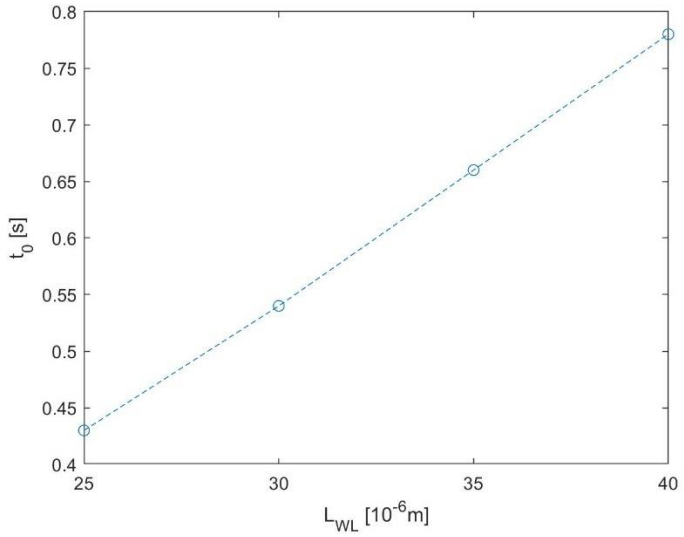


Figura 2.35 andamento di t_0 in funzione di L_{WL}

Per quanto riguarda l'impatto dello scaling sulle prestazioni del dispositivo, è stata osservata una riduzione di t_0 di un fattore 1.8 (figura 2.35), passando da una dimensione longitudinale del gate e dello spazio intercella di 40nm a 25nm (L_{WL}). La motivazione di questa variazione è da ricercare principalmente nella variazione di carica di lacune immagazzinata nella stringa all'equilibrio, piuttosto che nel processo di generazione di lacune ai bordi della stringa. Infatti, sebbene nella prima fase del transitorio venga osservata una forte differenza rispetto al caso studiato precedentemente, in seguito la generazione di lacune diventa paragonabile.

Capitolo 3

Modello circuitale

La necessità di riprodurre le dinamiche osservate nelle simulazioni numeriche, in maniera semplice e senza l'ausilio di software specifici, ha richiesto la modellazione circuitale, potenzialmente utile alla progettazione di dispositivi più resistenti al DCP. Di seguito verrà spiegato l'approccio usato per riprodurre i meccanismi dominanti e successivamente saranno confrontati i risultati numerici con quelli circuitali.

3.1 Selezione dei Bipoli

Per riprodurre i risultati, è stato determinante associare ad ogni meccanismo fisico un elemento circuitale, dipendente esclusivamente da tensione e corrente. Il primo fenomeno valutato è stato l'immagazzinamento di carica nella stringa, proporzionale a una tensione continua. L'elemento che modella questo comportamento è il condensatore, la cui carica è legata alla tensione secondo la relazione 3.1, dove C è la capacità del condensatore. Idealmente, in assenza di altri elementi circuitali o parassiti, la carica presente sulle armature varia nel tempo seguendo fedelmente la variazione della differenza di potenziale ai capi del bipolo, invece nelle simulazioni è stata osservata una dinamica molto più lenta che ha richiesto l'introduzione di altri elementi che producessero un ritardo.

$$Q = CV \tag{3.1}$$

Per modellare i fenomeni di trasporto e generazione e ricombinazione è stata selezionata la resistenza, tramite il cui valore è possibile modificare la quantità di carica rilasciata dai condensatori nel tempo. Nel caso della resistenza (R), tensione (V) e corrente (I) sono legati secondo la relazione 3.2.

$$V = RI \tag{3.2}$$

3.2 Modellazione dei condensatori

Sebbene il dispositivo in analisi non presenti regioni con drogaggio di tipo p e nonostante le tensioni applicate siano sempre non negative, la presenza di carica negativa nel nitruro implica una forte concentrazione di lacune nella stringa per basse tensioni, dominante rispetto agli elettroni, come mostrato in figura (3.3). A partire da queste considerazioni, è stato ritenuto opportuno modellare separatamente i due tipi di carica.

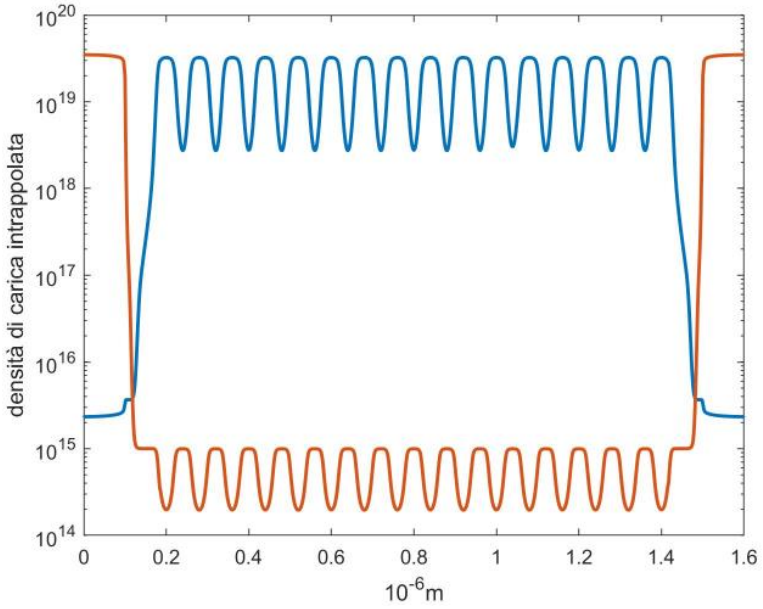


Figura 3.3 confronto delle densità di carica (cm^{-3}) all'interfaccia regione di canale-tunneling oxide. elettroni intrappolati (rosso), lacune intrappolate (blu)

Dalle simulazioni, è stata osservata una relazione tra carica e tensione fortemente non lineare, di conseguenza è stata abbandonata l'idea di modellare l'immagazzinamento di carica tramite una capacità costante. In aggiunta, la necessità di modellare anche l'occupazione delle trappole e l'effetto elettrostatico della carica di programmazione ha richiesto un approccio più specifico. Inizialmente, per il calcolo dei portatori liberi, è stata considerata la possibilità di impiegare un procedimento semi-analitico basato sulla statistica *maxwell-boltzmann*, ma la presenza del livello di fermi in *banda di conduzione* (Fig. 3.4) ha richiesto l'uso della statistica *Fermi-Dirac* (3.5, 3.6), in cui Ef indica il livello di fermi, mentre Ec ed Ev corrispondono rispettivamente al

limite inferiore della *banda di conduzione* e al limite superiore della *banda di valenza*. Nc ed Nv indicano la densità degli stati per gli elettroni e per le lacune.

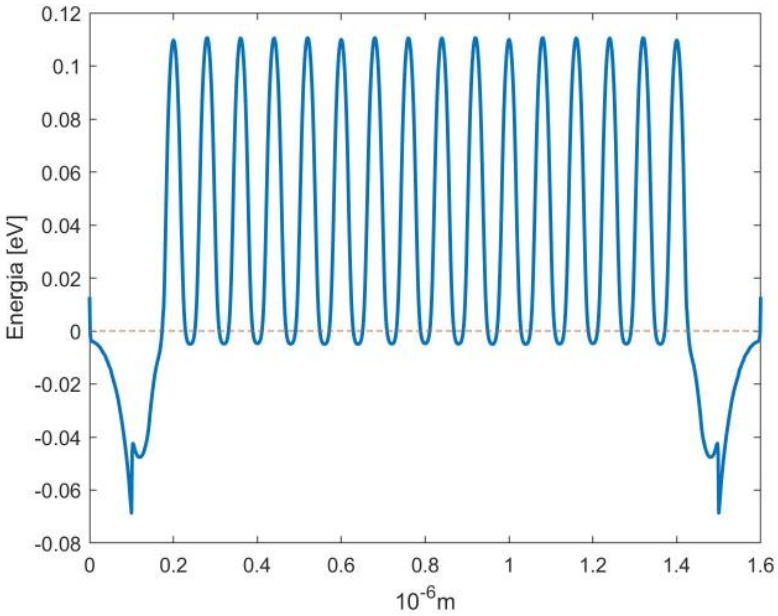


Figura 3.4 Profilo longitudinale della banda di conduzione (blu), livello di fermi (giallo)

$$n = Nc \cdot \frac{2}{\sqrt{\pi}} \int_0^{+\infty} \frac{\sqrt{x}dx}{1+e^{x-(Ef-Ec)/KT}} \tag{3.5}$$

$$p = Nv \cdot \frac{2}{\sqrt{\pi}} \int_{-\infty}^{Ev} \frac{\sqrt{x}dx}{1+e^{x-(Ev-Ef)/KT}} \tag{3.6}$$

La complessità introdotta dall'utilizzo delle equazioni 3.5 e 3.6 e la necessità di conoscere l'andamento del potenziale nella stringa hanno avuto come risultato la decisione di risolvere l'equazione di *Poisson* in tutta la struttura, implementando un codice su Matlab. Di seguito viene riportata la forma in coordinate cartesiane (3.7) e in coordinate cilindriche (3.8), con la derivata rispetto all'angolo pari a zero, visto che la struttura è cilindrica.

$$\frac{\partial^2 \varphi}{\partial x^2} + \frac{\partial^2 \varphi}{\partial y^2} + \frac{\partial^2 \varphi}{\partial z^2} = -\frac{\rho(\varphi)}{\varepsilon} \tag{3.7}$$

$$\frac{1}{r}\frac{\partial\varphi}{\partial r}+\frac{\partial^2\varphi}{\partial r^2}+\frac{\partial^2\varphi}{\partial z^2}=-\frac{\rho(\varphi)}{\varepsilon} \tag{3.8}$$

Per la corretta risoluzione di questa equazione alle derivate parziali, tramite il metodo delle differenze finite, si consideri dapprima il membro a sinistra dell'uguale. Seguendo questo procedimento, è necessario discretizzare le derivate, partendo dallo *sviluppo di taylor* di una funzione generica, come mostrato di seguito:

$$f(x+\Delta x)\cong f(x)+\ f'(x)\cdot\Delta x+\frac{f''(x)}{2}\Delta x^2 \tag{3.9}$$

$$f(x-\Delta x)\cong f(x)-\ f'(x)\cdot\Delta x+\frac{f''(x)}{2}\Delta x^2 \tag{3.10}$$

Trascurando il termine di secondo ordine, si ottiene la forma discretizzata della derivata (3.11).

$$f'(x)\cong(f(x\mp\Delta x)-f(x))/\Delta x \tag{3.11}$$

Mentre, sommando (3.10) e (3.9) si ottiene l'espressione della derivata seconda (3.12)

$$f''(x)=(f(x+\Delta x)-2f(x)+f(x-\Delta x))/\Delta x^2 \tag{3.12}$$

Per quanto riguarda il secondo membro, dove $\rho(\varphi)$ rappresenta la densità di carica volumetrica ed ε la permettività elettrica, nelle regioni prive di carica è pari a zero, nelle regioni di nitruro sottostanti ai gate delle celle ha un valore costante, invece nelle regioni drogate, assumendo una ionizzazione totale degli atomi droganti, ha un valore costante e uno è legato al potenziale tramite la statistica Fermi-Dirac. Come accennato in precedenza, la valutazione dell'integrale (3.5) e (3.6) risulta molto complessa, quindi, si è ricorsi ad una forma approssimativa (3.13, 3.14) che garantisce un errore massimo relativo di circa 0.4% rispetto alla formula tradizionale dell'integrale di Fermi-dirac di ordine 1/2. ($x=\frac{Ef-Ec}{KT}$ per gli elettroni, $\frac{Ev-Ef}{KT}$ per le lacune).

$$F_{\frac{1}{2}}(x) \cong (e^{-x} + \frac{3\sqrt{\pi}}{4} v^{-\frac{3}{8}})^{-1} \tag{3.13}$$

$$v = x^4 + 50 + 33.6x(1 - 0.68\exp(-0.17(x+1)^2)) \tag{3.14}$$

Trattandosi di un sistema studiato all'equilibrio termodinamico e con un livello di fermi costante in tutto il polisilicio, Ef è stato posto uguale a zero. A questo punto, aspettandosi un diagramma a bande discontinuo, a causa dei materiali coinvolti, non è stato possibile legare il potenziale al livello di fermi intrinseco (Ei), bensì è stato scelto il vacuum level, E0. Di conseguenza si è potuto legare il potenziale al diagramma a bande del sistema tramite le relazioni (3.15, 3.16). In tabella 3.17 vengono specificati i valori delle grandezze usate, in parte raccolti dalla letteratura, in parte estratti dalle simulazioni effettuate su sdevice. θ è l'affinità elettronica, $\Delta\varphi$ è la differenza tra la *workfunction* del contatto di gate e dell'ossido, φ_{ref} è la distanza tra E0 e il livello di fermi intrinseco.

$$-q(\varphi - \varphi_{ref}) = \mathrm{E0} = \mathrm{Ec} - \theta \tag{3.15}$$

$$\mathrm{Ev} = \mathrm{Ec} - \mathrm{Eg} \tag{3.16}$$

ϵ_0	$8.85 \cdot 10^{-14} \frac{F}{cm}$
ϵ_{siO2}	$3.9 \cdot \epsilon_0$
ϵ_{si3N4}	$7.5 \cdot \epsilon_0$
ϵ_{si}	$11.7 \cdot \epsilon_0$
$\Delta\varphi$	0.1663 V
Nc	$2.86 \cdot 10^{19} cm^{-3}$
Nv	$3.11 \cdot 10^{19} cm^{-3}$
Eg	1.1241 eV
φ_{ref}	4.6132 eV
θ	4.05 eV

Tabella 3.17 parametri fisici di simulazione

Un ulteriore passo fondamentale, prima della risoluzione effettiva dell'equazione, corrisponde alla definizione delle condizioni al contorno. Per quanto riguarda il potenziale all'interfaccia ossido-gate, è stata implementata la relazione (3.18), per le wordline il potenziale è stato imposto tramite la condizione di dirichlet (3.19), mentre, il potenziale all'interfaccia polisilicio-bitline è stato calcolato secondo (3.20), dove phi0 è il valore di potenziale calcolato dal principio di neutralità di carica in prossimità del contatto (3.21). Per le pareti esterne della struttura, idealmente a contatto col vuoto, è stata usata la condizione al contorno di *Neumann* discretizzata (3.22), dove φ_i è il potenziale all'interfaccia, invece per le interfacce dielettrico-semiconduttore e dielettrico-dielettrico è stata applicata la legge di Gauss (3.23).

$$\varphi_{interfaccia} = \varphi_{gate} - \Delta\varphi \tag{3.18}$$

$$\varphi_{gate} = \varphi_{WL} \tag{3.19}$$

$$\varphi_{Si} = \varphi_{BL} + phi0 \tag{3.20}$$

$$n = p + Nd \tag{3.21}$$

$$\frac{\varphi_i - \varphi_{i-1}}{\Delta x} = 0 \rightarrow \varphi_i = \varphi_{i-1} \tag{3.22}$$

$$\epsilon_1 F_1 = \epsilon_2 F_1 \qquad 23)$$

Per ottenere un buon grado di precisione nei calcoli, nel minor tempo di esecuzione possibile, la struttura descritta nel capitolo 1 è stata riprodotta con una griglia da 16261 nodi. Il passo internodale lungo la direzione longitudinale della stringa (Δz) è stato impostato pari a 10nm, invece, lo step nella direzione radiale (Δr) è stato posto uguale a 1nm.

$$(\varphi_{i,j+1} - \varphi_{i,j})/\,\Delta z + \,(\varphi_{i+1,j} - 2\varphi_{i,j} + \varphi_{i-1,j})/\Delta r^2 + (\varphi_{i+1,j} - \varphi_{i,j})/\,(i \cdot \Delta r) = -\,\frac{\rho(\varphi)}{\varepsilon} \quad (3.24)$$

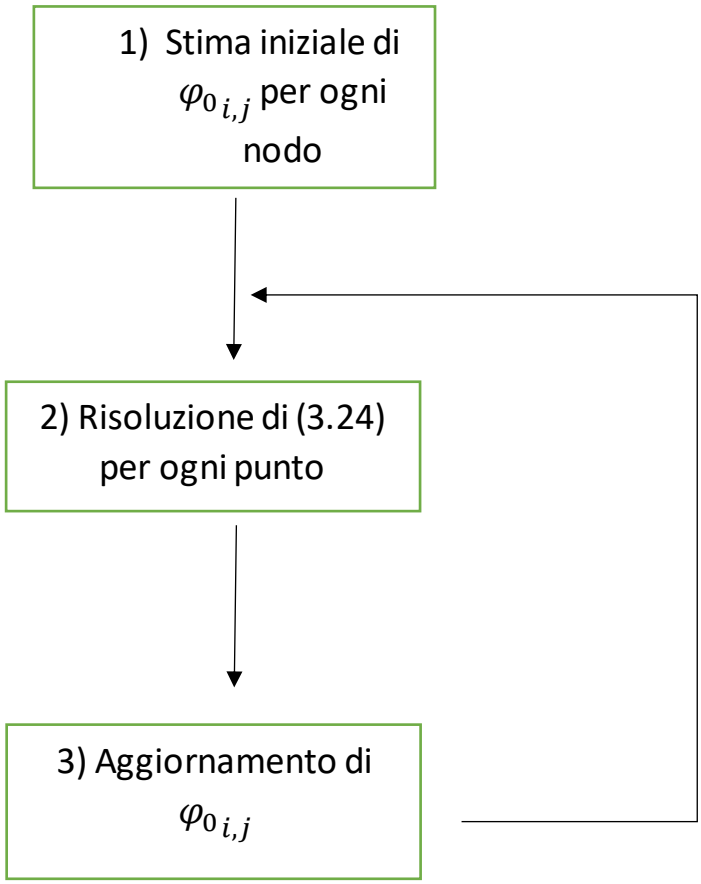
Partendo dall'equazione di Poisson discretizzata (3.24), è possibile isolare il termine $\varphi_{i,j}$ e si nota subito che il suo valore dipende dal potenziale dei quattro nodi ad esso adiacenti, anch'essi ignoti, se non all'interfaccia con i contatti. Di conseguenza si deduce che è necessario impostare un sistema per la risoluzione di queste equazioni. In aggiunta, la forma approssimativa dell'integrale di fermi dirac di ordine ½ (3.13) rende il sistema non lineare, quindi è necessario procedere nel seguente modo, al fine di ottenere un sistema lineare di facile risoluzione.

Dapprima, è necessario linearizzare la funzione (3.13), tramite il polinomio di taylor, intorno a un punto generico (φ_0) che si ritiene possa essere una buona stima della soluzione finale, come mostrato in (3.25).

$$\rho(\varphi_{i,j}) = f(\varphi_{i,j}) \cong f\left(\varphi_{0\,i,j}\right) + \,f'\left(\varphi_{0\,i,j}\right) \cdot (\varphi_{i,j} - \varphi_{0\,i,j}) \qquad (3.25)$$

Successivamente, dopo aver stimato una soluzione generica per ogni punto della griglia, sostituendo (3.25) al membro destro di (3.24), si può procedere con il metodo iterativo di newton per il calcolo della soluzione in due modi diversi. Nel primo, risolvendo per l'incognita $\varphi_{i,j}$, si ha una prima approssimazione della soluzione $f\left(\varphi_{i,j}\right)$, ottenuta con una sola iterazione. Proseguendo ugualmente per tutti i punti, si ottiene un insieme di soluzioni dovute a una sola iterazione. A questo punto si linearizza nuovamente la funzione ponendo $\varphi_{0\,i,j} = \varphi_{i,j}$ e si ripete

il processo fino al raggiungimento della precisione desiderata, come mostrato nello schema 3.26. La convergenza è assicurata a patto che le derivate non si annullino, circostanza che non si verifica nel nostro caso.



(3.26)

Nel secondo caso, invece di risolvere un'equazione per volta, si procede impostando una matrice dei coefficienti del sistema di equazioni lineari che deriva dalla sostituzione di (3.25) in (3.24) e generando un vettore colonna di termini noti, tramite le stime iniziali di $\varphi_{0_{i,j}}$. In seguito, si risolve per il vettore colonna ignoto e si ottiene la soluzione, per tutti i punti, per un'iterazione. Linearizzando ogni equazione intorno alla soluzione appena trovata e ripetendo il processo, si ottiene la soluzione finale. Il secondo metodo converge molto più velocemente del primo, ma si possono incontrare problemi di stabilità numerica, in quanto, per alcuni valori

di $\varphi_{0_{i,j}}$ la matrice dei coefficienti può essere singolare. Sono stati implementati entrambi i metodi, con esito positivo, ma si è deciso di procedere facendo affidamento sul primo. Per tener conto della carica intrappolata negli stati in banda proibita, è stato introdotto un coefficiente trapBC che modifica la densità degli stati in banda di conduzione, e un coefficiente trapBV che agisce sulla densità degli stati in banda di valenza. Sebbene questa scelta implichi una concentrazione errata di carica libera, permette tuttavia la previsione della carica totale nella stringa. Infatti, poiché questi calcoli sono destinati alla modellazione dei condensatori, non si potrebbe in ogni caso distinguere la posizione di questa carica nel diagramma a bande.

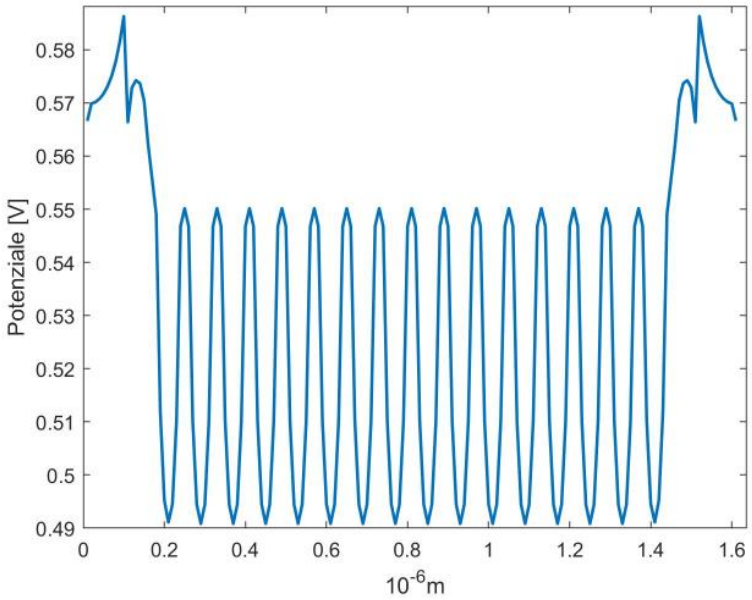


Figura 3.27 Profilo longitudinale del potenziale calcolato tramite Matlab (Vgate = 8V, Vbl/Vsl = 0V)

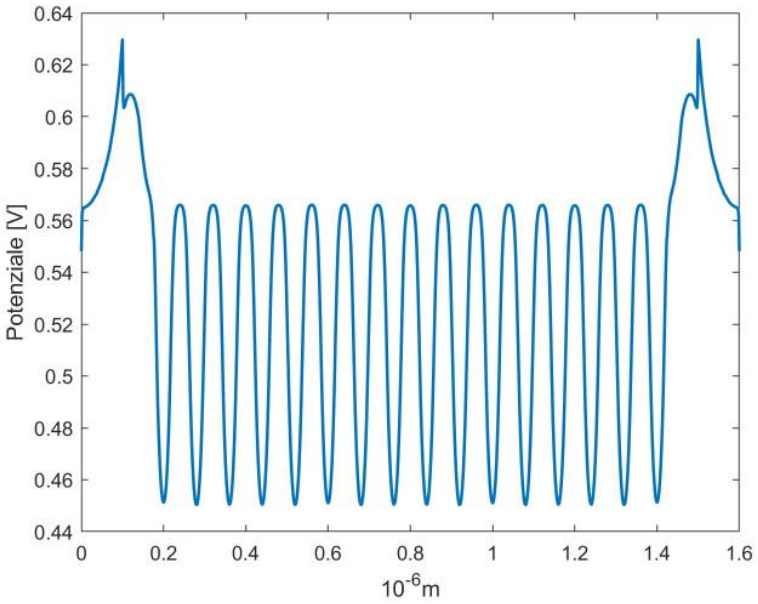


Figura 3.28 Profilo longitudinale del potenziale calcolato tramite Sdevice (Vgate = 8V, Vbl/Vsl = 0V)

A testimonianza della validità dei risultati, in figura 3.27 viene riportato il profilo longitudinale del potenziale della stringa, all'interfaccia tra tunneling oxide e regione di canale, mentre i risultati delle simulazioni effettuate su sdevice sono illustrati tramite il grafico (3.28). In (3.29) si mostrano i valori assunti nella sezione radiale in corrispondenza del centro del gate.

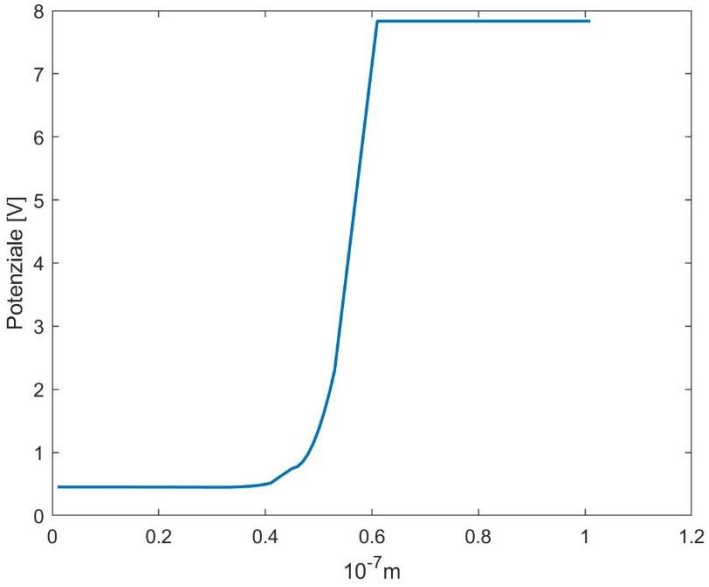


Figura 3.29 Profilo radiale del potenziale calcolato tramite Matlab (Vgate = 8V, Vbl/Vsl = 0V)

Per calcolare i condensatori da usare nel circuito, è stata fatta variare la tensione delle wordlines da 8V a 0V con passo 0.1V (Vstep). Per ogni valore è stata calcolata la carica di lacune ed elettroni (Qn e Qh), corrispondente alla sezione indicata in figura 3.30 e, a partire da questi risultati, sono state definite in maniera differenziale la capacità di elettroni (Cn) e la capacità di lacune (Ch), come indicato dalle relazioni (3.31, 3.32).

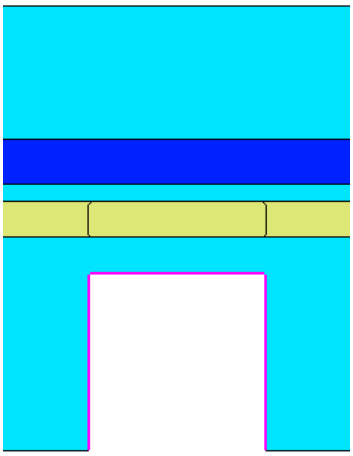


Figura 3.30 superificie di integrazione della densità di carica

$$Cn\,(V) = \frac{Qn(V) - Qn(V - Vstep)}{Vstep} \tag{3.31}$$

$$Ch\,(V) = \frac{Qh(V) - Qh(V - Vstep)}{Vstep} \tag{3.32}$$

L'andamento di Cn e Ch, in funzione della tensione, viene riportato in (3.32) e (3.33).

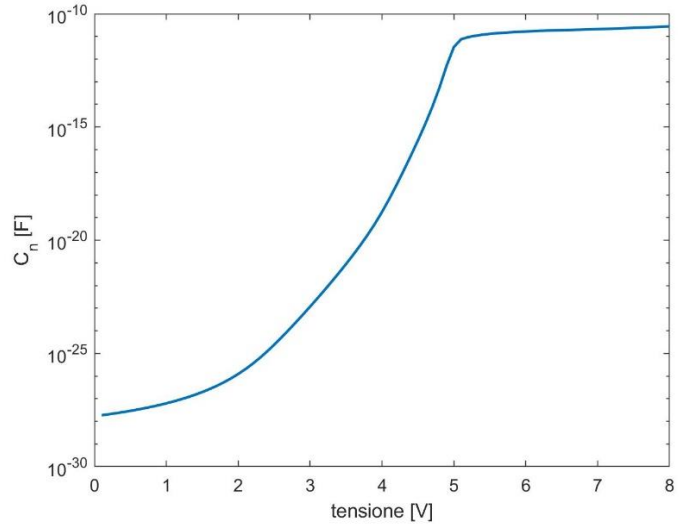


Figura 3.32 Cn in funzione della tensione applicata alla wordline

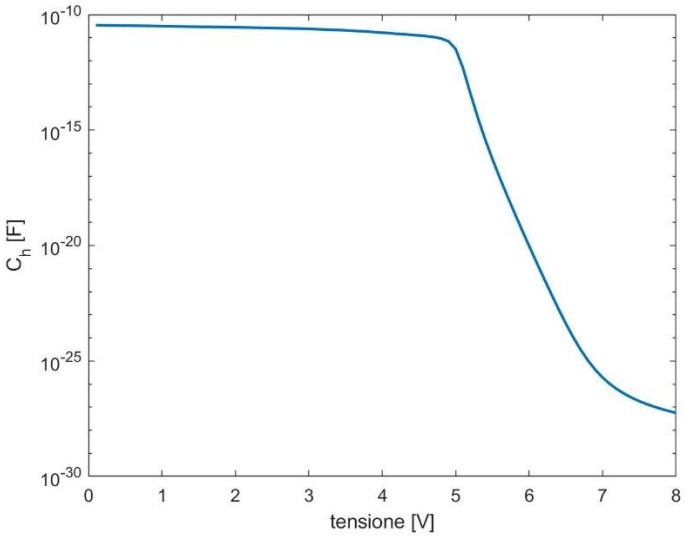


Figura 3.32 Ch in funzione della tensione applicata alla wordline

3.3 Modellazione resistenza

Per comprendere come modellare la scarica dei condensatori, è necessario analizzare l'andamento della banda di conduzione del centro della stringa simulato da sdevice e riportato

in figura 3.33. La simulazione è stata effettuata facendo variare la tensione delle wordlines da 8V a 0V in $10^{-15}s$, poiché solo modellando le dinamiche più veloci è possibile avere un'indicazione sul valore di potenziale per tempi maggiori. Si riconoscono, come spiegato precedentemente, tre regimi. Il primo è dipendente esclusivamente dagli elettroni, poiché la concentrazione di lacune è trascurabile, e dopo una rapida scarica di elettroni liberi e di difetti prossimi alla banda di conduzione, con brevi costanti di tempo, è caratterizzato da una pendenza logaritmica che indica il lento rilascio di carica da parte degli stati accettori. Il secondo, riconoscibile tramite il rapido cambio di pendenza, descrive l'arrivo di una forte concentrazione di lacune che interagiscono con gli elettroni, favorendo processi di ricombinazione. Nel terzo, il potenziale è sostenuto solo dalle lacune e gli unici processi di immissione e generazione di lacune avvengono ai bordi. Considerato che l'interazione tra i due portatori di carica è rilevante solo per una parte del transitorio, si è ritenuto opportuno modellare tramite due circuiti separati il comportamento di queste cariche.

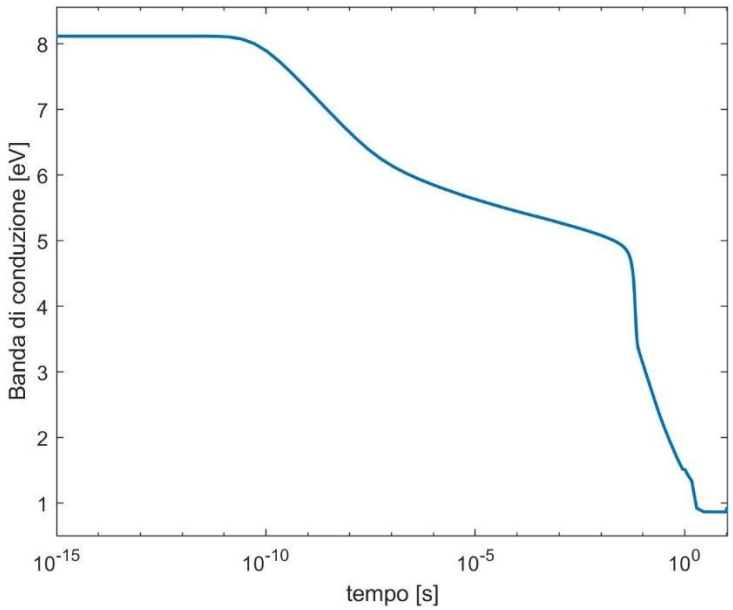


Figura 3.33 andamento della banda di conduzione al centro della struttura in funzione del tempo

Come svolto per i condensatori, è stata stimata la carica di ogni cella per ogni valore di tensione, in seguito è stata definita una densità media di carica volumetrica (n_m, h_m) e tramite le formule 3.34 e 3.35 è stata calcolata la resistenza Rn, dove Δr e Δz indicano la dimensione della cella.

$$\rho_n = \frac{1}{q A_n n_m} \tag{3.34}$$

$$\text{Rn} = \rho_n \cdot \frac{\Delta z}{2\pi \Delta r} \tag{3.35}$$

Il parametro An, seppur dimensionalmente assimilabile alla mobilità, non descrive solo i fenomeni di trasporto di portatori liberi, bensì anche la cessione di carica da parte dei difetti. Infatti, esso assume il valore della mobilità solo per tempi molto brevi (tensioni alte), per descrivere la fuoriuscita degli elettroni dal canale e per i difetti con costanti di tempo brevissime. In seguito, viene calibrato per riprodurre la pendenza logaritmica osservata nel transitorio.

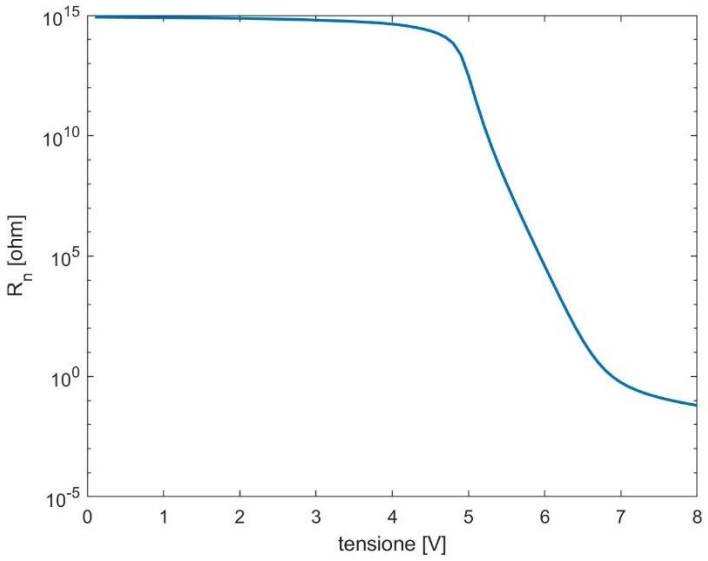


Figura 3.36 Rn in funzione della tensione applicata alla wordline

Il grafico 3.36 mostra la resistenza associata alla scarica di elettroni di una cella, in funzione della tensione. Si è proceduto in maniera analoga per le lacune. In questo caso, però, il parametro Ap acquisisce il valore della mobilità per basse tensioni, quando la densità di lacune è molto elevata. Per tensioni più alte, invece, la trasmissione di lacune, da una cella più estema a una più interna, è affidata al TAT e all'emissione PF, quindi, Ap è stato calibrato di conseguenza.

$$\rho_h = \frac{1}{qA_p h_m} \tag{3.37}$$

$$\mathrm{Rp} = \rho_h \cdot \frac{\Delta z}{2\pi\Delta r} \tag{3.38}$$

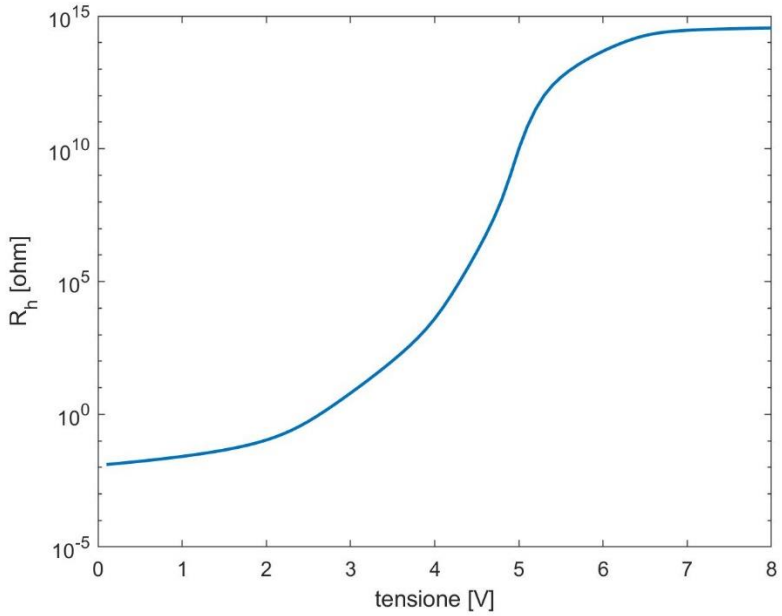


Figura 3.39 Rh in funzione della tensione applicata alla wordline

Per replicare la dinamica studiata in (3.33), non è stato necessario modellare la variazione di carica di elettroni degli estremi del dispositivo, in quanto avviene nell'arco di alcuni picosecondi e non influisce sui processi del centro della stringa. Invece, per quanto riguarda le lacune, è stato necessario introdurre una resistenza (Rg) con l'obiettivo di controllare la corrente

che riproduce i fenomeni di generazione ai bordi. A tal fine, questo elemento è stato dimensionato cercando di emulare, tramite la corrente, l'andamento nel tempo del contributo di generazione di lacune calcolato da *sdevice*. In particolare, per tempi superiori a 0.1s, istante in cui la concentrazione di lacune è dominante rispetto agli elettroni in tutta la struttura. A questo punto, per dimensionare Rg è sufficiente tenere a mente che G rappresenta il numero di lacune generate al secondo per unità di volume, quindi, moltiplicando per la carica elementare (q) e integrando rispetto al volume (ΔV) in cui avviene questo processo, si ottiene dimensionalmente una corrente (3.40). Sostituendo il risultato in (3.41), si ottiene la relazione 3.42, dove V indica il potenziale della cella di memoria più esterna.

$$I = G \cdot q \cdot \Delta V \tag{3.40}$$

$$I = \frac{V}{R_g} \tag{3.41}$$

$$Rg = \frac{V}{G \cdot q \cdot \Delta V} \tag{3.42}$$

Poiché l'andamento della concentrazione di lacune al centro della struttura (figura 3.43) è approssimabile a un processo di immissione esponenziale, si è deciso di usare un valore costante di Rg ($1.7 \cdot 10^8 ohm$), che prevedesse con maggiore accuratezza i valori finali di G, più determinanti per la stima della durata del transitorio.

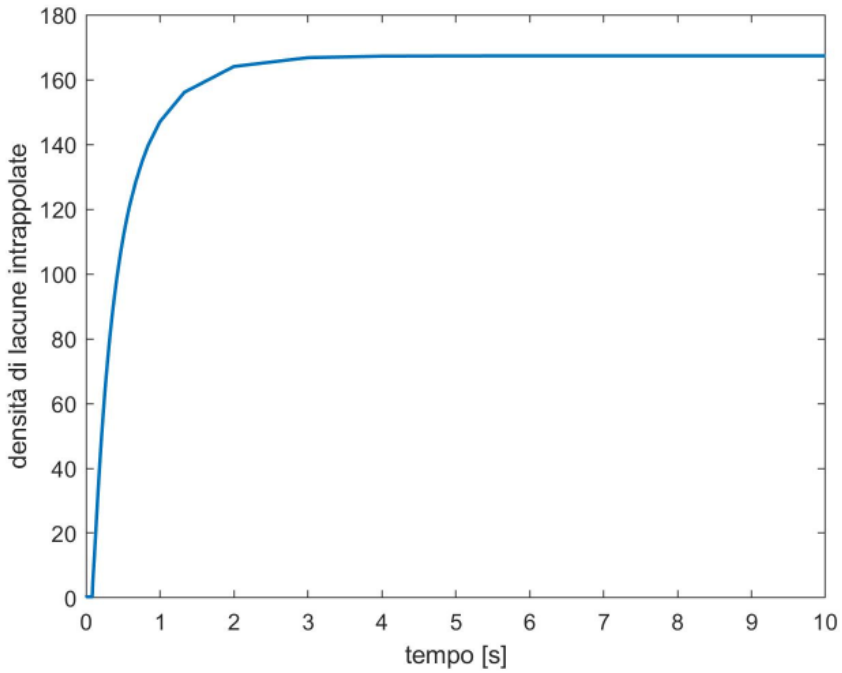


Figura 3.43 densità di carica di lacune intrappolate nella cella centrale

Sebbene questo approccio sembri mancare di generalità, la scelta di questo specifico valore di resistenza ha mostrato un buon grado di precisione per la stima della dinamica, sia al variare del numero di celle, sia al variare della dimensione longitudinale dei gate e dello spazio intercella.

3.4 Simulazione circuitale

Definiti i bipoli per gli elettroni e per le lacune, si può procedere con la valutazione della dinamica del sistema, tramite modello circuitale. Come esposto in precedenza, considerato che il potenziale è sostenuto inizialmente solo dagli elettroni e alla fine solo dalle lacune, si è deciso di ricorrere ad un circuito che modella solo la carica di elettroni per il calcolo del potenziale nella prima parte del transitorio, mentre, per la seconda parte è stato usato un circuito con carica di sole lacune. Nonostante questa scelta, entrambi i circuiti vengono svolti per tutta la durata della simulazione e vengono fatti interagire nella fase in cui il potenziale dipende da entrambi i tipi di carica. In particolare, l'interazione consiste nella sottrazione del numero di lacune da

quello di elettroni per simulare la ricombinazione di elettroni in seguito all'arrivo delle lacune generate ai bordi.

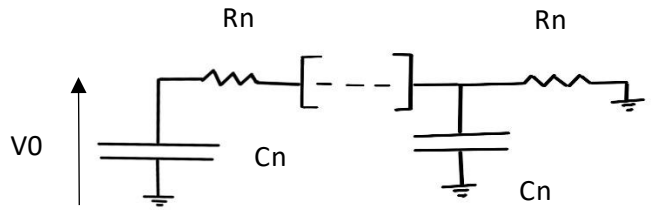


Figura 3.44 circuito di elettroni

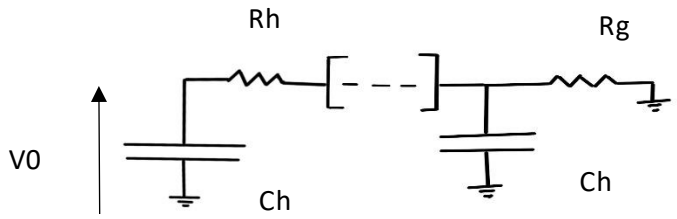


Figura 3.45 circuito di lacune

In figura 3.44 e 3.45 vengono raffigurati i circuiti descritti in precedenza. Ogni condensatore, che rappresenta una cella, ha una resistenza ad esso associata che ne modella la scarica. Entrambi gli elementi dipendono dalla tensione applicata ai capi della capacità (V0), secondo le dipendenze illustrate precedentemente. È d'obbligo precisare che queste dipendenze erano state valutate per ogni tensione di gate da 8V a 0V in condizioni statiche, mentre in tutto il corso della simulazione circuitale i gates sono fissi a 0V e la stringa ha un potenziale variabile nel tempo.

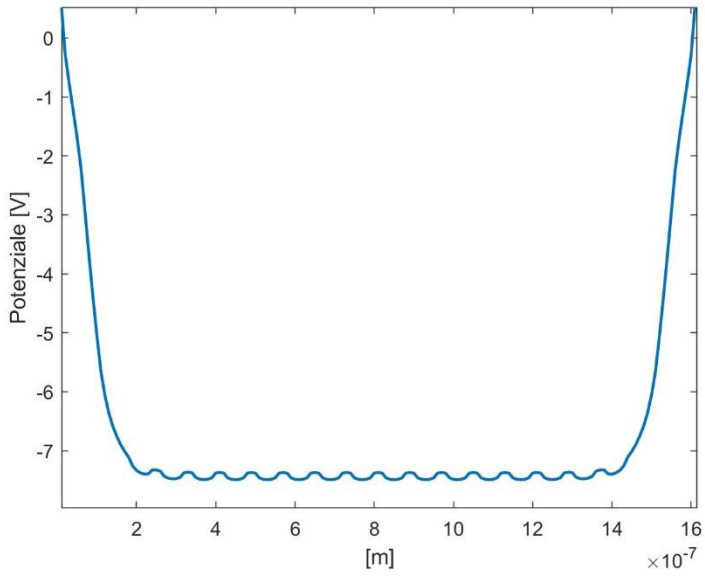


Figura 3.46 Potenziale della stringa in seguito a uno scalino di -8V

Per comprendere perché è possibile usare le medesime dipendenze, è utile indagare l'effetto dell'abbassamento del gate da 8V a 0V, in un tempo infinitesimo, sul potenziale della stringa. Alla rapida variazione della tensione di gate, non corrisponde una variazione di carica nel dispositivo, quindi avviene uno shift del diagramma a bande pari al gradino di tensione applicato. Infatti, imponendo la stessa carica nella stringa e fissando tutti i contatti a 0V, è possibile calcolare, tramite l'equazione di Poisson, il profilo del potenziale all'interfaccia della regione di canale, come mostrato in figura 3.46. Osservando questo grafico, si nota subito che, poiché la tensione sulla stringa ha subito uno shift di 8V, la differenza di potenziale ai capi del condensatore è ancora la stessa, ma di segno opposto, invece, la carica è rimasta costante, quindi la capacità non è variata, ma si è modificato solo il segno di V0. Ripetendo questo ragionamento per tutte le tensioni tra 8V e 0V si deduce che Cn(|V0|) = Cn(V_{gate}).

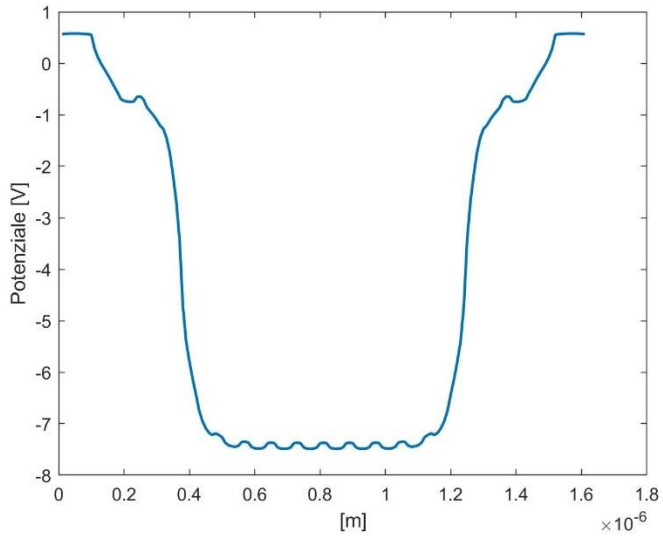


Figura 3.47 Potenziale della stringa in seguito a uno scalino di -8V, con cariche nelle prime due celle pari al valore di equilibrio

Il limite di questa analisi è la scarsa precisione nella stima del potenziale delle celle esterne tramite la relazione Qn = CnV0. Infatti, in quella sezione del dispositivo, la carica varia molto più rapidamente rispetto al centro della stringa e quindi nel momento in cui raggiunge una concentrazione di carica prossima all'equilibrio, prevede una tensione di circa 0V. In figura 3.47, viene invece mostrato che, ponendo la carica nelle prime due celle pari alla carica all'equilibrio ($V_{WL} = 0$V), mentre la carica nelle altre celle è pari al valore iniziale, il potenziale delle due celle, calcolato tramite l'equazione di Poisson, è ancora distante circa 1V dal valore finale. Si può estendere lo stesso ragionamento alla posizione del livello di fermi di lacune e di elettroni che risulta di più facile impiego. Questo limite è di ostacolo solo per il circuito di elettroni, infatti, quello di lacune, intorno all'equilibrio, modella una variazione di potenziale costante su tutta la stringa. Fatte queste precisazioni, si specifica che, per ragioni di simmetria, viene risolta solo metà struttura e il calcolo dell'evoluzione libera del sistema ha luogo in seguito all'abbassamento, tramite un gradino di -8V, della tensione di gate. Quindi tutti i condensatori hanno un valore iniziale del potenziale della stringa pari a V0 -8V.

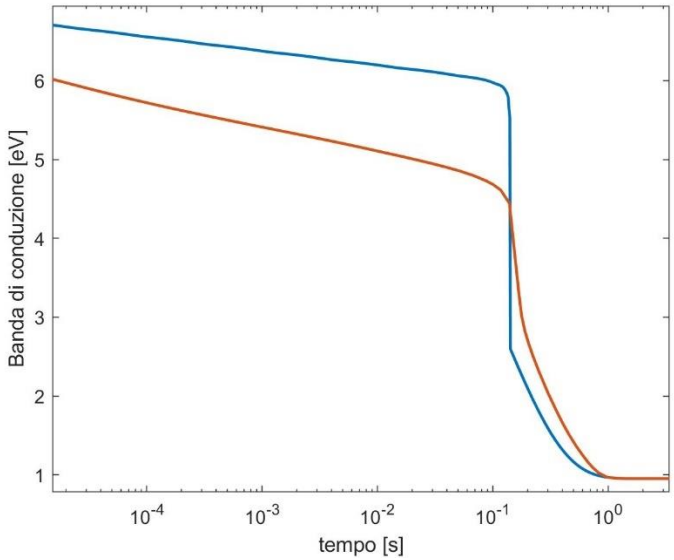


Figura 3.48 Andamento della banda di conduzione simulato tramite il circuito (blu) e tramite sdevice (rosso), per 64 celle

In figura 3.48, viene messo a confronto il risultato calcolato tramite questo modello con le simulazioni TCAD. La scelta di un valore costante di resistenza (Rg) non permette una perfetta corrispondenza per tempi prossimi a 0.1s, dato che G assume valori superiori che giustificano la pendenza maggiore della curva in rosso, ma l'accuratezza aumenta avvicinandosi all'equilibrio. In tabella 3.49 viene confrontato il tempo t0 in cui la stringa raggiunge il 10% del valore finale, anche per strutture con un numero maggiore di celle.

Numero celle	t_0 circuito	t_0 sdevice
64	0.6s	0.77s
128	1.18s	1.49s
256	2.36s	3.09s

Tabella 3.49 confronto dei tempi di equilibrio

3.5 Modello semi-analitico

Partendo dai risultati ottenuti numericamente, è stato possibile semplificare ulteriormente il modello, al fine di offrire un riferimento per il calcolo degli elementi coinvolti nella simulazione. Per quanto riguarda la capacità Cn, si nota una regione di plateau, il cui inizio indica orientativamente la tensione di soglia. Al variare della carica nel trapping layer, questa regione si estende o si restringe proporzionalmente, come mostrato in figura 3.50, in cui nP assume valori -10^{19}, -1.5×10^{19}, -2×10^{19}, -2.5×10^{19} $[cm^{-3}]$. Una modifica analoga ha luogo per la resistenza Rn.

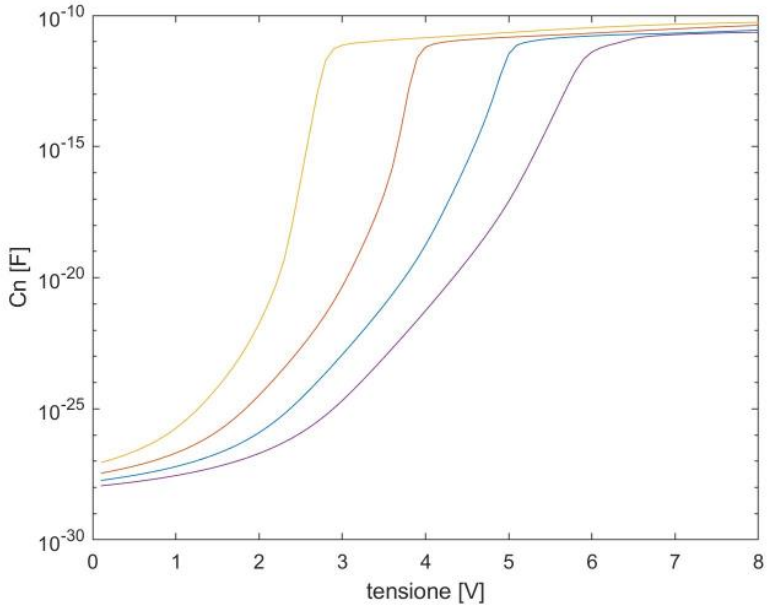


Figura 3.50 Cn in funzione della tensione al variare di nP: -10^{19} (giallo), -1.5×10^{19} (arancione), -2×10^{19} (blu), -2.5×10^{19} (viola)

Poiché per alte tensioni i valori sono molto simili per tutte le curve, così come la pendenza, il valore di Cn è stato considerato indipendente dalla carica di programmazione ed è stato definito in maniera esponenziale, come espresso secondo la relazione 3.51 e mostrato in 3.52, dove C0 indica il valore di capacità ottenuto prolungando la regione con pendenza minima fino a 0V e assume valore -7.5×10^{-13} F. Hn è pari a 2.22V ed è un coefficiente estrapolato numericamente per approssimare la pendenza osservata in figura 3.50.

$$Cn_a(V) = C0e^{V/Hn} \tag{3.51}$$

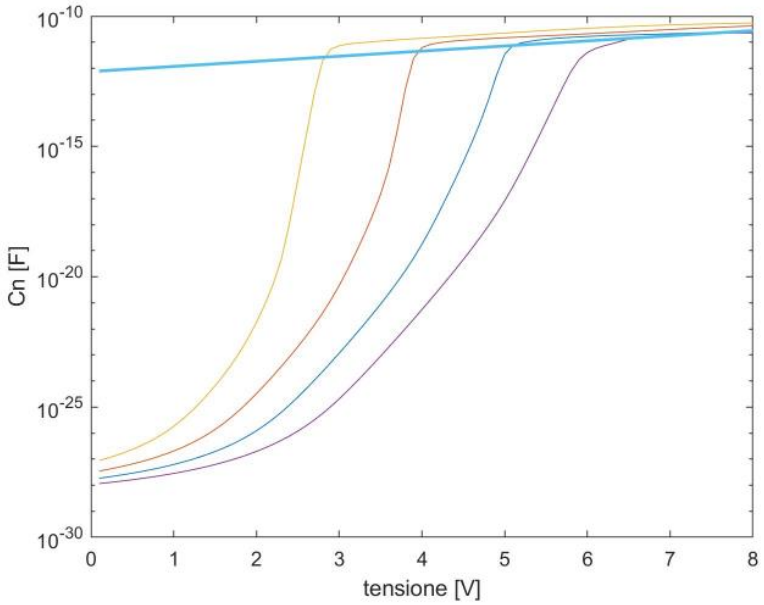


Figura 3.52 Cn in funzione della tensione al variare di nP: -10^{19} (giallo), -1.5×10^{19}(arancione), -2×10^{19}(blu), -2.5×10^{19} (viola), Cn_a (azzurro)

Per quanto riguarda Rn, si è proceduto associando alla regione sopra-soglia un valore costante di resistenza (R0 = 50 ohm), mentre per simulare la diminuzione di conducibilità per tensioni minori di V_{th}, e la scarica delle trappole è stato introdotto un termine esponenziale. L'introduzione di quest'ultimo termine permette una maggiore libertà e una più semplice calibrazione, necessaria in quanto non è nota a priori la distribuzione e la concentrazione di stati in banda proibita. Nella relazione 3.53 viene espresso quantitativamente il valore di Rn e si notano i parametri A e B, tramite i quali dimensionare Rn. In questo caso V_{th} è circa 5.5V, A è pari a $10^3 ohm$ e B vale 0.13V. Questo approccio empirico ha il fine di stimare l'andamento della banda di conduzione nella fase iniziale, ma poiché il raggiungimento dell'equilibrio dipende quasi esclusivamente dalle lacune, un errore di dimensionamento di questi parametri non influirà sulla parte finale della curva.

$$Rn(V) = R0 + Ae^{-(V-V_{th})/B} \tag{3.53}$$

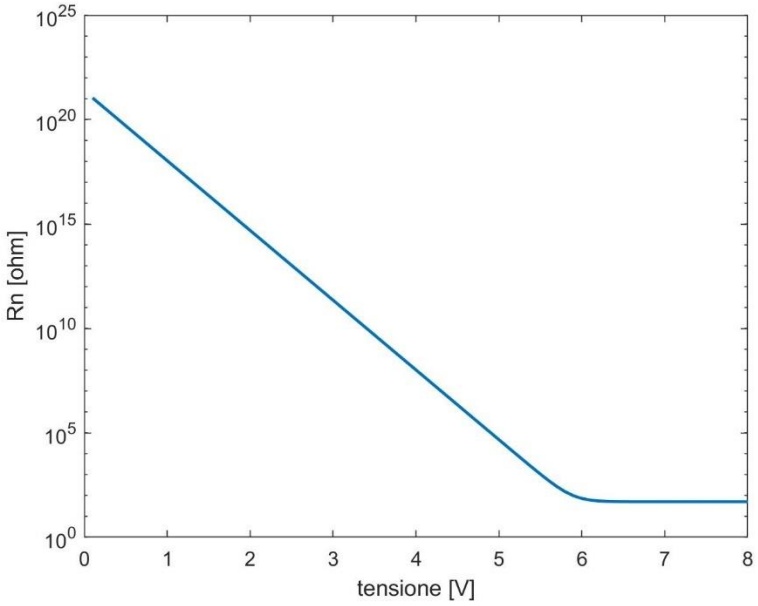


Figura 3.52 Rn in funzione della tensione

Per quanto riguarda Ch, è stato usato un procedimento analogo a quello impiegato per Cn ottenendo la relazione 3.53. Ch(0V) indica il valore calcolato numericamente, corrispondente al valore di capacità massimo di valore circa $3.53 \times 10^{-11} F$. Hh è pari a 3.12V.

$$Ch(V) = Ch(0V)e^{-V/Hh} \tag{3.53}$$

Per il calcolo delle resistenze associate alle lacune, bisogna effettuare una distinzione tra resistenza Rh e resistenza Rg. La prima modella il trasporto di lacune tra le celle e appena la concentrazione di lacune diventa dominante, il suo valore tende a zero, quindi, per basse tensioni, la caduta di tensione su di esse è trascurabile e le capacità di lacune risultano in parallelo. La seconda controlla l'immissione di lacune nella stringa. Per quanto riguarda la prima, si è deciso di trascurare la dipendenza dalla tensione e porla a un valore costante di 0.1 ohm. Così facendo, si assume che non ci siano ritardi nella propagazione di lacune verso il centro dovuti alla altre celle, che non corrisponde alla realtà, però dato che esse diventano dominanti nella stringa per tempi sufficientemente minori rispetto al tempo di equilibrio, è possibile effettuare questa approssimazione. Con questa decisione si semplifica di molto il calcolo, in quanto è necessario modellare solo l'iniezione di lacune nella parte finale del

transitorio, quando tutte le capacità sono in parallelo. Come osservato in precedenza, i fenomeni di generazione legati alle trappole e al campo elettrico sono dominanti quindi il dimensionamento di Rg è stato effettuato a partire dall'espressione del tasso di generazione SRH, con una modifica sul tempo di vita legata al campo elettrico (F), secondo le dipendenze esposte nel capitolo 2. Per semplicità, il numeratore è stato pari a ni^2, per il tempo di vita τ_0 è stato selezionato il valore massimo suggerito dal software sentaurus, pari a $10^{-5}s$, ed è stato considerato il contributo dovuto solo agli stati a metà banda proibita, in quanto dominanti. La densità delle trappole a metà banda (Nt = $7.75\times 10^{15}cm^{-3}$) è stata calcolata supponendo la stessa distribuzione integrata nelle simulazioni TCAD, ossia un andamento esponenziale con concentrazione totale di $5 \times 10^{19}cm^{-3}$ e una costante di energia pari a 110meV, che moltiplica l'esponente. La densità di elettroni è stata ritenuta trascurabile in quanto inferiore rispetto alle lacune. Di seguito, nell'espressione 3.54, si mostra come è stata legata la densità di lacune ai parametri circuitali e come è stato dimensionato il parametro di riduzione di τ_0, γ_{PF}.

$$G = Nt\frac{ni^2\gamma_{PF}}{\tau_0 p} \tag{3.54a}$$

$$p = \frac{Ch(V)}{q\Delta z\Delta r 2\pi} \tag{3.54b}$$

$$\gamma_{PF} = \frac{1}{A^2}(1 + (A-1)e^A) - 1/2 \tag{3.54c}$$

$$A = \sqrt{\frac{q}{KT}\sqrt{\frac{qF}{\varepsilon_{poly}\pi}}} \tag{3.54d}$$

In figura 3.55 viene mostrato l'andamento di G in funzione del potenziale, da 8V a 0V, pari a un campo elettrico costante da $10^6\ Vcm^{-1}$ a $0Vcm^{-1}$, localizzato tra la prima cella della stringa e la regione del contatto di bitline, che corrisponde a un'estensione di 80nm. Si nota un andamento esponenziale, fatta eccezione per la parte delle basse tensioni, che causa una lieve dilatazione dei tempi, come era stato osservato anche nelle simulazioni precedenti sull'analisi del transitorio.

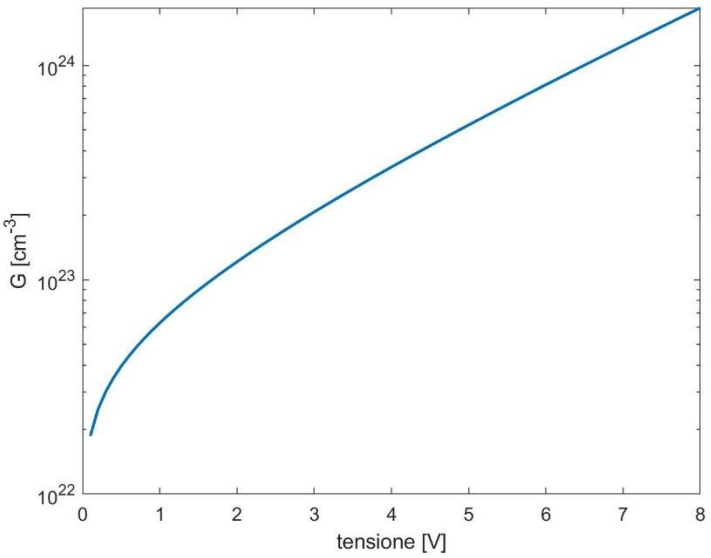


Figura 3.55 Generazione in funzione della tensione

Per dimensionare Rg in prima approssimazione, ricordando la relazione esposta nel capitolo 2 che lega generazione e corrente, è sufficiente calcolare il rapporto tra tensione (8V) e corrente (circa 16 nA per G = 10^{24}). Con questo valore si ottiene Rg = $5 \times 10^{8} ohm$, valore simile a quello stimato precedentemente, che in presenza di un condensatore di valore costante, assicura un andamento esponenziale. Per tener conto della dipendenza di Ch dalla tensione e al fine di simulare il cambio di pendenza per basse tensioni, il valore di Rg selezionato è stato $1.5 \times 10^{9} ohm$. Analogamente a quanto eseguito in precedenza, i due circuiti vengono risolti simultaneamente, ma poiché non è stata modellata la propagazione di lacune verso il centro, avendo posto un valore costante di Rh, si può produrre solo un grafico asintotico, risultante dall'intersezione dei valori di banda di conduzione calcolati separatamente. Per tempi minori dell'istante di intersezione, l'andamento è definito dal circuito di elettroni, in seguito da quello di lacune. Di seguito, in figura 3.56 viene confrontato l'andamento energetico della cella centrale calcolato a partire da queste espressioni con quello simulato tramite sdevice. In figura 3.57 e 3.58 si mettono invece a confronto i valori di banda di conduzione anche della prima cella.

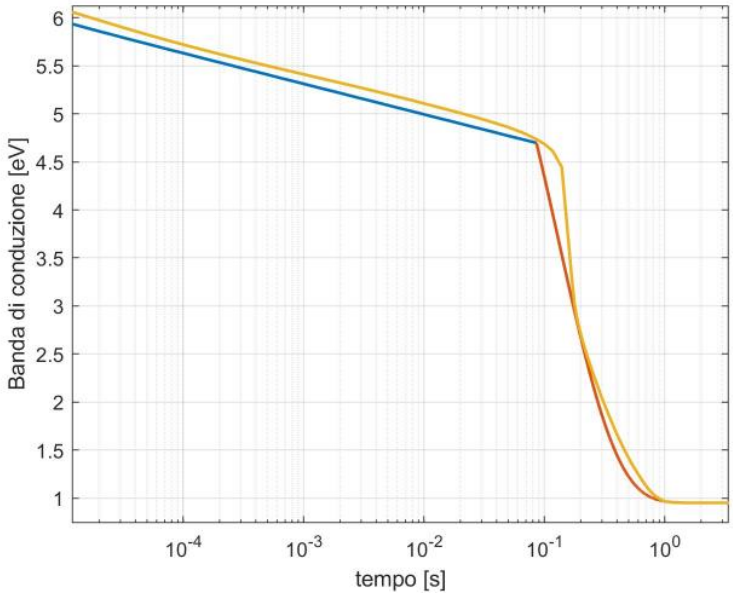


Figura 3.56 andamento della banda di conduzione calcolato dal modello (rosso e blu), simulato con sdevice (giallo)

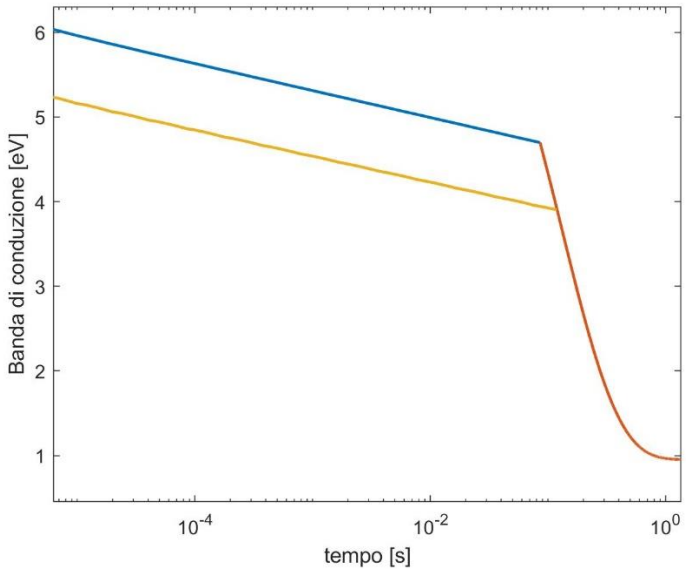


Figura 3.57 andamento della banda di conduzione della cella centrale (rosso e blu), della prima cella (giallo e rosso). Simulazione modello

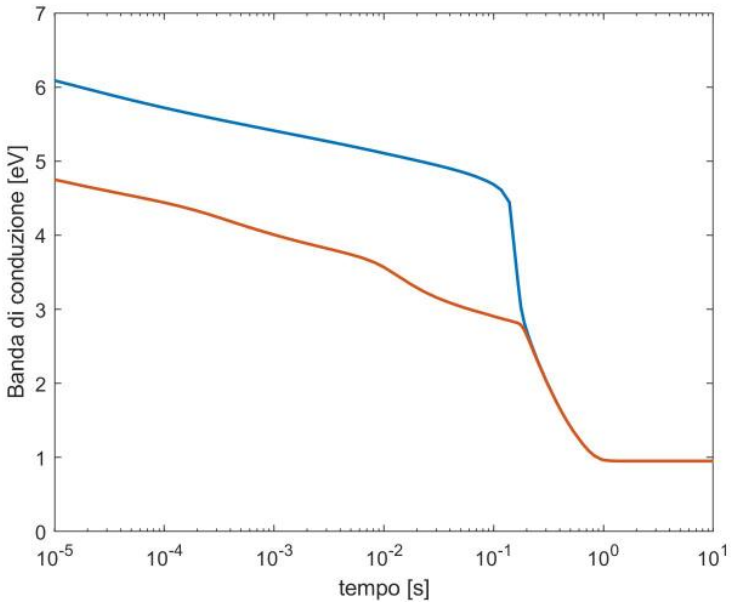


Figura 3.58 andamento della banda di conduzione della cella centrale (blu), della prima cella (rosso). Simulazione TCAD

Dal confronto della figura 3.58 e 3.57, si nota una maggior diminuzione della banda di conduzione della cella esterna, dovuto all'immissione di lacune. Nel modello circuitale, invece, le lacune hanno influenza su tutta la stringa solo quando raggiungono un valore dominante rispetto agli elettroni nella cella centrale, quindi, non si assiste a una variazione dovuta ad esse nelle periferie della stringa.

3.6 Modello semplificato

Per evitare di risolvere numericamente il circuito descritto in precedenza, e per avere una stima della durata della dinamica, è possibile effettuare qualche approssimazione. In particolare, partendo dall'assunto che il raggiungimento dell'equilibrio è dipendente quasi esclusivamente dalla generazione di lacune ai bordi, è possibile non modellare la prima parte del transitorio, in quanto la scarica degli stati accettori non influisce sull'arrivo di lacune. Inoltre, come osservato nelle simulazioni, per tempi sufficientemente minori rispetto al tempo di raggiungimento dell'equilibrio, il potenziale della stringa è costante. Questo implica che la caduta di tensione

su tutte le resistenze (tranne Rg) è trascurabile e che i condensatori possono essere considerati in parallelo. Poiché si è prossimi al valore di equilibrio, si può approssimare Ch con Ch(0V), visto l'andamento di Ch per basse tensioni. Con queste approssimazioni il sistema si riduce a un circuito con un condensatore di valore Ch(0V)·N/2 e una resistenza in serie pari a Rg (N indica il numero di celle). Di conseguenza si può stimare la costante di tempo del sistema ponendo τ = Ch(0V)·N/2 ·Rg. Inoltre, modellando Ch(0V) come un condensatore cilindrico con armature metalliche, e quindi imponendo una dipendenza lineare di C con L_{WL}, si può introdurre un fattore di scaling S (3.59), dove L_{ref} è pari a 40nm. Con questa ulteriore approssimazione, si ottiene la relazione 3.60.

$$S = \frac{L}{L_{ref}} \tag{3.59}$$

$$\tau = Ch \cdot Rg \cdot \frac{N}{2} \cdot S \tag{3.60}$$

Questo procedimento risulta molto utile al variare della dimensione longitudinale dei gate e della distanza intercella. Infatti, il diverso profilo del campo elettrico non permette di descrivere con precisione la curva, senza un'ulteriore calibrazione tramite le simulazioni numeriche. Però, è stato osservato che l'andamento del potenziale, per tempi prossimi all'equilibrio, è assimilabile ai casi analizzati precedentemente, di conseguenza è possibile usare il medesimo valore di Rg. In aggiunta poiché Rg modella i processi di generazione, è sufficiente agire su questo parametro per tener conto della variazione delle grandezze legate alle trappole. Questa approssimazione si rivela anche utile, in quanto non è più necessario conoscere l'occupazione degli stati donori durante tutta l'analisi, bensì solo all'equilibrio. Inoltre, poiché il livello di fermi può essere considerato coincidente con il limite superiore della banda di valenza, si può assumere un'occupazione totale degli stati donori. Con queste considerazioni si può esprimere l'ultima relazione che lega τ ai parametri descritti, ottenendo l'espressione 3.61. In tabella 3.62 vengono confrontati i valori di 2.3τ (valore per cui la l'esponenziale raggiunge il 90% del valore finale) con t_0 calcolato da sdevice.

$$\tau = Ch \cdot Rg \cdot \frac{N}{2} \cdot S \tag{3.61}$$

Numero celle	2.3τ	t_0
64	0.46s	0.77s
128	0.92s	1.49s
256	1.84s	3.09s

Tabella 3.62 tempi di equilibrio a confronto, con N variabile, $L_{WL} = 40nm$

In tabella 3.63 si confrontano i risultati al variare della geometria. Fatta eccezione per il valore iniziale per N = 64 e L_{WL} = 40nm, le dipendenze vengono rispettate con sufficiente precisione. Infatti, notiamo una dipendenza lineare con N e una diminuzione di un fattore 1.6 da 40nm a 25nm, tramite τ, mentre si ha un fattore 1.79 dalle simulazioni TCAD.

L_{WL} (nm)	2.3τ	t_0
40	0.46s	0.77s
35	0.40s	0.66s
30	0.35s	0.54s
25	0.29s	0.43s

Tabella 3.63 tempi di equilibrio a confronto, con L_{WL} variabile, N = 64

Conclusioni

L'obiettivo di questo lavoro è stata la comprensione del DCP in una struttura BiCS. Con l'aiuto delle simulazioni numeriche è stato possibile distinguere i fenomeni dominanti per il ritorno all'equilibrio della stringa programmata, in seguito al passaggio delle wordlines da una tensione di passaggio, a massa. A tal fine, è stata determinante l'introduzione delle trappole della regione di canale in polisilicio, infatti i risultati hanno supportato l'idea che i processi di generazione e ricombinazione, legati agli stati in banda proibita, fossero alla base della dinamica di ritorno all'equilibrio. In particolare, è stata dimostrata la centralità dei meccanismi di carica e scarica legati al campo elettrico, come il trap-assisted tunneling e l'emissione Poole-Frenkel. Come indicatore dell'andamento del sistema, è stata usata l'energia della banda di conduzione, tramite il quale è stato possibile distinguere tre regimi di ritorno all'equilibrio. Nel primo avviene un lento rilascio di carica di elettroni intrappolata nei difetti accettori, mentre ha luogo una generazione di lacune ai bordi della stringa, dove è localizzato il campo elettrico più intenso. La graduale propagazione di lacune verso il centro del dispositivo assume un ruolo dominante quando la concentrazione di carica positiva supera quella negativa. A questo corrisponde l'abbassamento improvviso del potenziale di tutta la stringa che caratterizza il secondo regime. Infine, la riduzione del campo elettrico laterale modula una produzione di lacune inferiore e una dinamica più lenta che porta all'equilibrio. Il passo successivo è stato comprendere l'impatto del numero di celle e dello scaling della dimensione longitudinale delle wordline e degli spazi intercella. È stata osservata una dipendenza lineare tra numero di celle e tempo di equilibrio, in quanto il processo dominante ha luogo ai bordi della struttura e non è quindi dipendente dal numero di celle. A causa di ciò avviene un'immissione di lacune invariata a fronte di un aumento proporzionale di carica richiesta per il raggiungimento dell'equilibrio. La miniaturizzazione della geometria ha introdotto forti variazioni solo nelle fasi iniziali del transitorio, in cui è stato notato un forte anticipo del secondo regime. Tuttavia, la fase finale, più rilevante per il raggiungimento dell'equilibrio, ha mostrato molte similitudini con il caso iniziale, indicando il coinvolgimento degli stessi meccanismi. I risultati hanno quindi espresso una dipendenza quasi lineare tra tempo e F, per motivazioni analoghe a quelle descritte per l'aumento del numero di celle. A partire da queste considerazioni, ci si è dedicati alla modellazione circuitale di quanto analizzato. È stato possibile descrivere l'andamento del sistema nel caso

iniziale e all'aumentare del numero di celle ed è stata offerta una soluzione per stimare gli effetti dello scaling. Lo studio e la modellazione di questo fenomeno rimangono ancora di interesse, vista la necessità di una comprensione approfondita delle dipendenze dai parametri di trappola.

Bibliografia

[1] C. Monzio Compagnoni, A. Goda, A. S. Spinelli, P. Feeley, A. L. Lacaita, and A. Visconti, "Reviewing the evolution of the nand flash technology," *Proceedings of the IEEE*, vol. 105, no. 9, pp. 1609–1633, 2017. 6, 7, 8, 14, 15

[2] C. M. Compagnoni, A. S. Spinelli, R. Gusmeroli, A. L. Lacaita, S. Beltrami, A. Ghetti, and A. Visconti, "First evidence for injection statistics accuracy limitations in nand flash constant-current fowler-nordheim programming," in *2007 IEEE International Electron Devices Meeting*, pp. 165–168, 2007. 9

[3] D. Ielmini, A. Spinelli, A. Lacaita, and A. Modelli, "A new two-trap tunneling model for the anomalous stress-induced leakage current (SILC) in Flash memories," *Microelectronic engineering*, vol. 59, no. 1, pp. 189–195, 2001. 12

[4] K. Prall and K. Parat, "25nm 64gb mlc nand technology and scaling challenges invited paper," in *2010 International Electron Devices Meeting*, pp. 5.2.1–5.2.4, 2010. 12

[5] A. Ghetti, C. Monzio Compagnoni, A. S. Spinelli, and A. Visconti, "Comprehensive analysis of random telegraph noise instability and its scaling in deca–nanometer flash memories," *IEEE Transactions on Electron Devices*, vol. 56, no. 8, pp. 1746–1752, 2009. 13

[6] Y. Nishi, *Advances in Non-volatile Memory and Storage Technology*, ch. Developments in 3D-NAND Flash technology. Woodhead Publishing Series in Electronic and Optical Materials, Elsevier Science, 2014. 16

[7] J. Jang, H. Kim, W. Cho, H. Cho, Jinho Kim, S. I. Shim, Younggoan, J. Jeong, B. Son, D. W. Kim, Kihyun, J. Shim, J. S. Lim, K. Kim, S. Y. Yi, J. Lim, D. Chung, H. Moon, Sungmin Hwang, J. Lee, Y. Son, U. Chung, and W. Lee, "Vertical cell array using tcat(terabit cell array transistor) technology for ultra high density nand flash memory," in *2009 Symposium on VLSI Technology*, pp. 192–193, 2009. 17

[8] T. Y. Chan, J. Chen, P. K. Ko, and C. Hu, "The impact of gate-induced drain leakage current on mosfet scaling," in *1987 International Electron Devices Meeting*, pp. 718–721, 1987. 19, 20

[9] Y. Fukuzumi, R. Katsumata, M. Kito, M. Kido, M. Sato, H. Tanaka, Y. Nagata, Y. Matsuoka, Y. Iwata, H. Aochi, and A. Nitayama, "Optimal integration and characteristics of vertical array devices for ultra-high density, bit-cost scalable flash memory," in *2007 IEEE International Electron Devices Meeting*, pp. 449–452, 2007. 21

[10] Tae-Sung Jung, Young-Joon Choi, Kang-Deog Suh, Byung-Hoon Suh, Jin-Ki Kim, Young-Ho Lim, Yong-Nam Koh, Jong-Wook Park, Ki-Jong Lee, Jung-Hoon Park, Kee-Tae Park, Jhang-Rae Kim, Jeong-Hyong Yi, and Hyung-Kyu Lim, "A 117-mm/sup 2/ 3.3-v only 128-mb multilevel nand flash memory for mass storage applications," *IEEE Journal of Solid-State Circuits*, vol. 31, no. 11, pp. 1575–1583, 1996. 21, 23

[11] M. Kang and Y. Kim, "Natural local self-boosting effect in 3d nand flash memory," *IEEE Electron Device Letters*, vol. 38, no. 9, pp. 1236–1239, 2017. 21, 22

[12] Y. Kim and M. Kang, "Down-coupling phenomenon of floating channel in 3d nand flash memory," *IEEE Electron Device Letters*, vol. 37, no. 12, pp. 1566–1569, 2016. 24

[13] G. Masetti, M. Severi, and S. Solmi, "Modeling of carrier mobility against carrier concentration in arsenic-, phosphorus-, and boron-doped silicon," *IEEE Transactions on Electron Devices*, vol. 30, no. 7, pp. 764–769, 1983. 30

[14] C. Sah, T. Ning, and L. Tschopp, "The scattering of electrons by surface oxide charges and by lattice vibrations at the silicon-silicon dioxide interface," *Surface Science*, vol. 32, no. 3, pp. 561 – 575, 1972. 30

[15] A. Hartstein, T. Ning, and A. Fowler, "Electron scattering in silicon inversion layers by oxide and surface roughness," *Surface Science*, vol. 58, no. 1, pp. 178 – 181, 1976. 30

[16] A. Mannara, A. S. Spinelli, A. L. Lacaita, and C. Monzio Compagnoni, "Current transport in polysilicon-channel gaa mosfets: A modeling perspective," in *ESSDERC 2019 - 49th European Solid-State Device Research Conference (ESSDERC)*, pp. 222–225, 2019. 32

[17] D. M. Kim, A. N. Khondker, S. S. Ahmed, and R. R. Shah, "Theory of conduction in polysilicon: Drift-diffusion approach in crystalline-amorphous-crystalline semiconductor system—part i: Small signal theory," *IEEE Transactions on Electron Devices*, vol. 31, no. 4, pp. 480–493, 1984. 32

[18] N. C. . Lu, L. Gerzberg, Chih-Yuan Lu, and J. D. Meindl, "Modeling and optimization of monolithic polycrystalline silicon resistors," *IEEE Transactions on Electron Devices*, vol. 28, no. 7, pp. 818–830, 1981. 32

[19] A. Bansal, B. C. Paul, and K. Roy, "An analytical fringe capacitance model for interconnects using conformal mapping," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 25, no. 12, pp. 2765–2774, 2006. 50