UNIVERSITÀ
DEGLI STUDI
DI PADOVA

DIPARTIMENTO
DI INGEGNERIA
DELL'INFORMAZIONE

DIPARTIMENTO DI INGEGNERIA DELL'INFORMAZIONE

CORSO DI LAUREA IN INGEGNERIA ELETTRONICA

"ANALISI DELLO STATO DELL'ARTE DELLE MEMORIE DALL'ESIGENZA ALLA NECESSITÀ"

"STATE OF THE ART OF MEMORIES FROM REQUEREMENT TO NECESSITY"

Relatore: Prof. Vogrig Daniele

Laureando: Leonte Cosmin

ANNO ACCADEMICO 2022 – 2023

Data di laurea 16 NOVEMBRE 2023

Sommario

RINGRAZIO:

Il docente relatore Daniele Vogrig, il quale di ha aiutato ed assistito durante la redazione di questo documento

La mia famiglia, la quale mi ha dato la possibilità di studiare e mi ha sempre sostenuto nel mio percorso

I miei compagni di università Nicolò, Luca, Zimmo, Giulia che hanno reso la vita universitaria più leggera e piacevole

Ai miei amici Ana, Alisa, Vittoria, Enrico, Giada, le quali mi hanno aiutato a prendere decisioni importanti e tirato su nei momenti più bui.

Un ringraziamento speciale va alla mia Segretaria Lucrezia, senza la quale probabilmente non sarei Dottore oggi

Grazie a tutti del sostegno

INTRODUZIONE

Negli ultimi anni, l'aumento della densità delle memorie, hanno permesso all'industria dei circuiti integrati di sviluppare circuiti sempre più complessi.
L'esigenza di memoria è un concetto fondamentale nel contesto dell'informatica, dei dispositivi elettronici e dell'archiviazione di dati. Questa necessità attuale può essere rappresentata da diversi fattori, come l'evoluzione tecnologica, l'aumento della frequenza e della quantità di dati da elaborare, l'aumento della complessità dei programmi e dei software, l'aumento della quantità di dati da mantenere ecc.…
Per ognuna di queste abbiamo uno o più aspetti delle memorie che vogliamo migliorare per essere meglio adattate alle nostre necessità.
Le caratteristiche di una memoria sono:

- Dimensione
- Parametri temporali
- Funzione
- Modo di accesso
- Architettura della porta di ingresso e di uscita
- Applicazioni

ARCHITETTURA A BLOCCHI DELLE MEMORIE

Questa è decisamente la soluzione più semplice e intuitiva per la creazione di una memoria.
Come illustrato nella Figura 1, il nucleo della memoria (parte in giallo) è dove vengono immagazzinati i dati. È composto celle elementari (storage cell) ordinate per righe e colonne. Per via del fatto che vogliamo memorie il più simmetriche possibili, le righe sono composte da più parole disposte una di fianco all'altra. Affinché il dispositivo funzioni correttamente, abbiamo bisogno di tre componenti aggiuntivi:

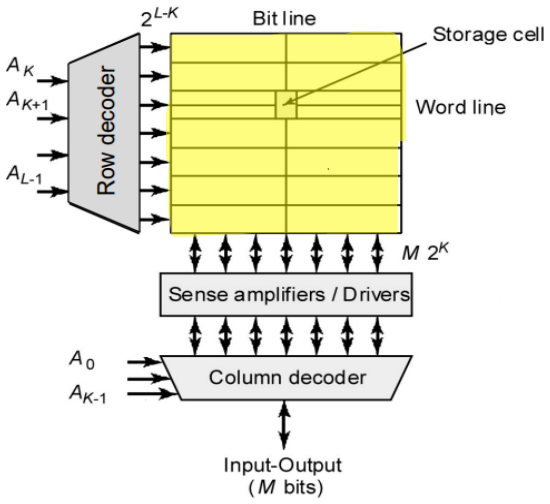


Figura 1: Architettura chip di memoria

- Row decoder (decoder di riga): riduce di un termine logaritmico il numero di segnali per selezionale una determinata Word line. Con n segnali possiamo determinare 2^n linee
- Sense amplifiers (amplificatore): si tende a diminuire l'escursione logica della bit line per ridurre sia il tempo di propagazione sia il consumo di potenza. Per utilizzare il dato all'esterno della memoria abbiamo bisogno di ripristinarlo ad un'escursione completa
- Column decoder (decoder di colonna: serve ad indirizzare la parola corretta verso i terminali di uscita (o di ingresso)

Questa configurazione è conveniente da usare per la realizzazione di memorie di dimensioni modeste. Volendo aumentare la dimensione, dobbiamo aggiungere una terza dimensione, come illustrato nella Figura 2. Esso è composto da P blocchi equivalenti alla Figura 1 uguali tra di loro e, essendo che abbiamo aggiunto una dimensione, dobbiamo aggiungere una Block address in modo da riuscire a identificare il blocco contenente l'informazione richiesta. In questo modo riusciamo ad aumentare la dimensione lasciando un tempo di propagazione minimo ed un consumo di potenza limitato (i blocchi non attivi restano in modalità di risparmio).

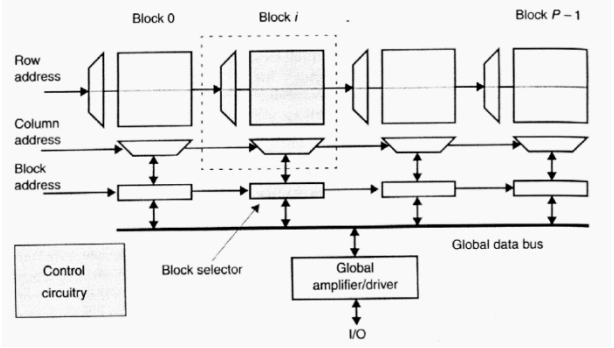


Figura 2: Chip memoria a tre dimensioni

TIPOLOGIE DI MEMORIA

Dalla precedente analisi delle celle di memoria, sembra immediato pensare che il problema principale per le grandi banchi di memoria sia quello di tenere ridotte le dimensioni del nucleo. Questo è sicuramente vero, ma non bisogna perdere di vista il fatto che, diminuendo le dimensioni delle celle, abbiamo una diminuzione delle prestazioni, come velocità e robustezza. Si distinguono tre tipologie di memorie: a sola lettura (ROM); a lettura-scrittura volatili; a lettura-scrittura non volatili (MVRWM).

- ROM: le Read Only Memories sono memorie scritte una volta sola e, come dice il nome, il loro contenuto può essere solamente letto. Questo avviene perché il contenuto della matrice dipende dal layout del circuito. Per quanto possano risultare strane, trovano una vasta applicazione in tutti qui campi dell'elettronica nelle quali non abbiamo bisogno di modificare il circuito ed i comportamenti (come gli elettrodomestici per esempio). Inoltre, hanno una minore complessità ed elevate prestazioni.

  Una particolare tipologia di questa categoria è la EPROM, una memoria di sola lettura programmabile, e cancellabile tramite appositi macchinari (il processo avviene fuori dal

circuito). Questa tipologia è stata quella che ha anticipato l'arrivo delle memorie non volatili di lettura e scrittura.

- MVRWM: le Non-Volatile Read-Write Memories hanno una struttura simile alle memorie ROM, la differenza sta nel fatto che il loro contenuto può essere riscritto più volte. Il nucleo della memoria è costruito da una matrice di celle contenenti transistor.
- RAM: questa tipologia di memoria, come la precedente, ha la possibilità di modificare il contenuto. La differenza sta nel fatto che è una memoria volatile, cioè, una volta che l'alimentazione viene a mancare, il contenuto della cella si cancella.

RANDOM ACCESS MEMORY

Nei sistemi digitali, il termine RAM viene adoperato per indicare una memoria con capacità sia di lettura che di scrittura. Questa tipologia viene usata quando abbiamo bisogno di modificare il suo contenuto frequentemente. Essa è infatti ampiamente utilizzata come memoria temporanea nei computer, ma hanno bisogno di essere costantemente collegate all'alimentazione affinché possano funzionare, in quanto sono delle memorie volatili.
Quando si avvia un programma, il processore invia un comando per recuperare tali programmi dalla memoria secondaria (non volatile). Il processore colloca i programmi nella RAM o nel contatore digitale, temporaneamente mentre lavora con loro in modo che possa accedere alle informazioni in modo più rapido e semplice. Perciò le prestazioni della memoria influenza anche quelle del dispositivo.
La realizzazione della memoria RAM parte dall'esigenza di avere tempi di scrittura e di lettura circa uguali. La memorizzazione delle informazioni in tali celle è basata sul principio di retroazione positiva o sulla carica immagazzinata all'interno di una capacità. Il problema di questi circuiti è dato dall'eccessiva occupazione di area. A seconda del tipo di approccio utilizzato per immagazzinare l'informazione all'interno di una cella è possibile catalogare le memorie RAM in due: statiche (SRAM) e dinamiche (DRAM).

MVRWM

Le memorie non volatili a letture-scrittura rappresentano una classe di dispositivi di archiviazione avanzati che hanno rivoluzionato il panorama dell'informatica. A differenza della tradizionale RAM, che è volubile e richiede alimentazione costante per mantenere i dati, le NVRWM consentono la conservazione permanente delle informazioni anche quando il sistema viene spento. Questo attributo fondamentale le ha rese un elemento chiave in una vasta gamma di applicazioni, dalla conservazione dei dati sensibili in dispositivi mobili all'accelerazione delle operazioni di avvio in computer di fascia alta. Altri settori in cui hanno trovato una vasta applicazione sono: l'elettronica di consumo, l'automotive, l'aerospaziale ecc.…. Nei computer,

vengono utilizzate per archiviare il BIOS (Basic Input/Output System) e altre informazioni di configurazione essenziali che devono sopravvivere ai riavvii. Inoltre, le NVRWM stanno diventando sempre più importanti nei sistemi di archiviazione di nuova generazione, in cui la combinazione di velocità, affidabilità e capacità di conservazione dei dati a lungo termine è fondamentale.

La principale categoria è la memoria FLASH, la quale offre una combinazione di velocità, affidabilità e basso consumo energetico ed è quindi ideale per applicazioni che richiedono una rapida lettura e scrittura di dati. Questa tipologia è ampiamente utilizzata in dispositivi di archiviazione, come SSD (Solid State Drive) e schede USB nella configurazione Flash NAND. Tuttavia, le NVRWM basate su memoria flash possono essere soggette a un numero limitato di cicli di scrittura, il che può influire sulla loro durata nel lungo termine.

STATIC RANDOM ACCESS MEMORY

Viene anche utilizzata come memoria cache per i via dei suoi bassi tempi d'accesso. Al contrario, è una memoria costosa e richiede un consumo elettrico elevato.
Il modello generale di una cella di memoria SRAM è rappresentato in Figura 3.

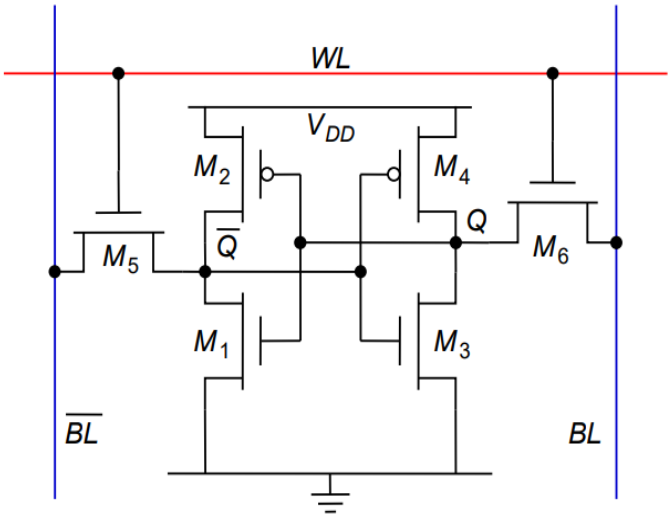


Figura 3: Cella SRAM 6T

È basata sulla tecnologia CMOS, infatti è un circuito bistabile con 2 inverter accoppiati come in un flip-flop e 2 transistor per l'accesso. Notiamo la presenza di due bit line per cella perché tale cella può fornire sia il valore logico memorizzato sia il suo negato. Questa configurazione non è strettamente necessaria, ma, così facendo, si migliora la robustezza e i margini di rumore. Notiamo subito che richiede un'area considerevole in una realizzazione integrata in quanto abbiamo bisogno di sei transistor per ogni cella.
Analizziamo il funzionamento della cella per le operazioni di lettura e scrittura.
Per l'operazione di lettura, le Bit line vengono precaricate ad una tensione pari a $V_{DD}/2$ in modo da aumentare le prestazioni e diminuire il consumo elettrico (si riduce l'escursione di tensione). Il ciclo di lettura inizia quando la Word line commuta al valore logico '1', il quale abilita i transistor M5-M6 con un certo ritardo dato dal tempo di propagazione lungo la linea. Le Bit line in corrispondenza dei due nodi si caricheranno o scaricheranno a seconda che nel nodo sia memorizzato un '1' o uno '0'. Per migliorare le prestazioni del circuito si può aggiungere un sense amplifier in modo tale da completare la carica della Bit line alla minima variazione della tensione. Tra il nodo a potenziale nullo (quello contenente lo '0' logico) e la BL ad una tensione di precarica, si genera una corrente che scorre dalla Bit line alla capacità interna del nodo. Questo passaggio si ferma quando la tensione della capacità del nodo interno eguaglia quella della Bit line, portando ad un effetto distruttivo del dato immagazzinato. Per evitare ciò, bisogna dimensionare i transistor interni in maniera tale da avere una tensione di soglia maggiore della tensione, in maniera tale da non accendere i transistor. Da questo punto

di vista, non è possibile diminuire le dimensioni della cella sotto una determinata soglia senza compromettere il comportamento del dispositivo.
Nel caso, invece, si voglie eseguire un'operazione di scrittura, le BL sono utilizzate per imporre il valore voluto all'interno della cella. In questa era di rapido sviluppo dei dispositivi portatili, abbiamo bisogno di sistemi che utilizzino la batteria in modo efficiente. Essendo che le celle di memoria SRAM non richiedono un refresh (ma mantengono il dato finché sono collegate all'alimentazione) sono preferibili alle memorie DRAM (illustrate nel Capitolo successivo) in questo tipo di applicazioni.
Il consumo di energia totale è pari a:

$$E_{Tot} = E_{Switching} + E_{Leakage}$$

$E_{Leakage}$ rappresenta l'energia statica consumata dovuta alla perdita di corrente della cella. Quando la Word line è ad un valore logico '0', i transistor di accesso M5 e M6 disconnettono la cella dalle Bit line. I 2 inverter accoppiati nella cella continuano a rinforzare il dato contenuto finché è collegata all'alimentazione V_{DD}. La corrente assorbita in questa modalità da V_{DD} è definita corrente di alimentazione.
$E_{switching}$ rappresenta l'energia consumata durante le operazioni di lettura o scrittura.

MIGLIORAMENTO DELLA SRAM 6T

Per ridurre il consumo di energia della cella a sei transistor, possiamo usare due tecniche combinate: gated V_{DD} e MTCMOS (Multi Threshold CMOS) [1]

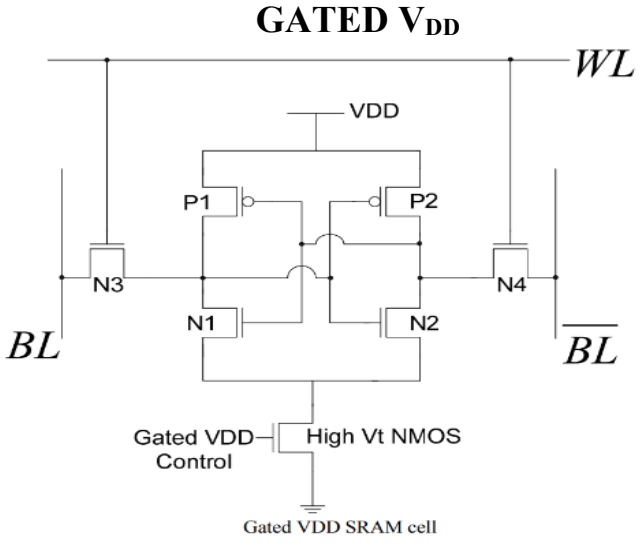


Figura 4: Cella SRAM 6T gated V_{DD}

Questa tecnica introduce un meccanismo di controllo che riesce a disattivare la sorgente di tensione quando la cella SRAM non è utilizzata, eliminando lo "spreco" di corrente.
Nella Figura 4 possiamo osservare la stessa configurazione precedente con l'aggiunta di un NMOS ad alta tensione di soglia Vt, mentre gli atri transistor sono modellati in modo tale che

abbiano una bassa tensione di soglia. Il segnale Gated VDD Control ha la funzione di accendere o spegnere la cella. Questa configurazione possiede un ritardo ed un consumo energetico ottimale, ma incrementa l'area di integrazione richiesta

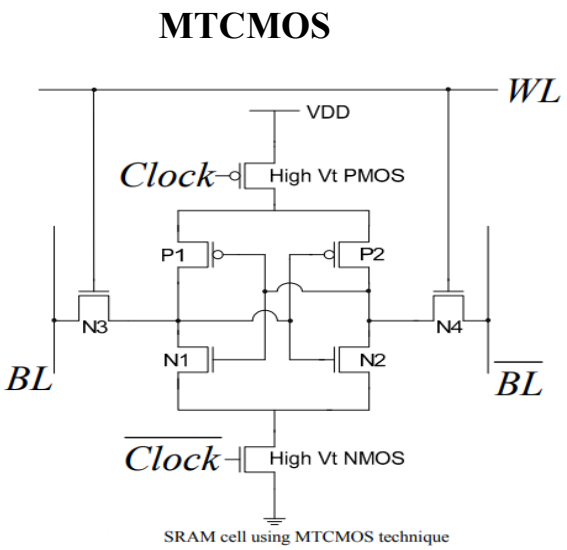


Figura 5: Cella SRAM 6T MTCMOS

Un dispositivo con una tensione di soglia Vt elevata diminuisce la perdita di corrente, ma è più lento. Al contrario, un Vt basso è decisamente più veloce ma ha un consumo di potenza statico alto. Questa tecnica implementa due sleep transistors (un PMOS nella rete di pull-up ed un NMOS nella rete di pull-down) con una tensione di soglia elevata. Lo stato della cella è dettato dal segnale di Clock. Con il loro aiuto siamo in grado di creare un percorso virtuale per alimentare il circuito. Il resto dei transistor ha una Vt bassa per mantenere le prestazioni del circuito elevate.

Quando i sleep transistors sono spenti, l'alimentazione ed il ground vengono disconnessi dalla cella. I due bit '0' e '1', posti ai lati della cella, creano un percorso di alimentazione virtuale caricandosi tramite i pass transistors degli inverter. Questo percorso ha un potenziale minore rispetto a V_{DD}, ma è abbastanza per rafforzare il dato memorizzato. Con il tempo, tale percorso virtuale tende ad attenuarsi, rischiando di far sparire il dato. Per evitare questo comportamento distruttivo, bisogna attivare i sleep transistor prima che questo accada, in modo da restaurare il bit.

RISULTATI E DISCUSSIONE

Gli schemi dei circuiti sono progettati e simulati con tecnologia a 90 nm [1] con V_{DD}=1,8V utilizzando Cadence Virtuoso. I sono dimensionati in modo tale da avere un ritardo uguale per le reti di pull-up e pull-down. I transistor NMOS hanno dimensioni di 120 nm, mentre i transistor PMOS hanno dimensioni di 300nm. I sleep transistors nella cella basata su MTCMOS sono entrambi dimensionati a 120 nm per ottenere ritardi e consumi energetici ottimali.

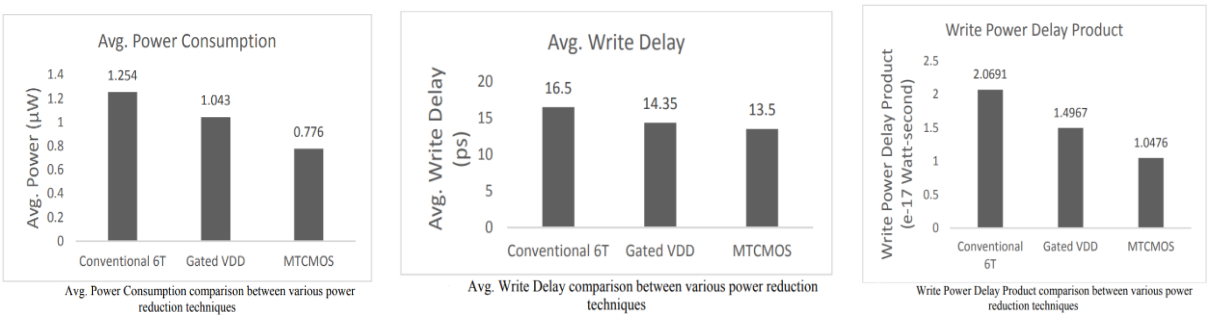


Figura 6: Confronto caratteristiche delle soluzioni

Nella Figura 6 possiamo osserviamo che le 2 tecniche migliorano la velocità e riducono il consumo delle celle, ma bisogna ricordare che per la loro realizzazione abbiamo aumentato il numero dei transistors, il che si traduce in un aumento dell'area occupata.

Miglioramenti 6T	**Avg. Power**	**Avg. Write Delay**	**Avg. Power Delay Product**
GATED V_{DD}	16.8%	13.03%	27.66%
MTCMOS	38.1%	18.18%	49.37%

Tabella 1

Nella Tabella 1 sono riportati i miglioramenti in percentuale rispetto alla cella SRAM 6T originale. MTCMOS è sicuramente la tecnica che offre prestazioni migliori, a discapito dell'occupazione di area. Essa ha bisogno di un transistor in più per cella rispetto al GATED V_{DD}.

CELLA SRAM A TRE TRANSISTOR

La memoria cache è di fondamentale importanza per mantenere le prestazioni di un sistema elevate. Essa viene usata dal processore per deporre programmi e dati di uso frequente, così da averli disponibili più velocemente senza andare a cercali nella memoria principale (più lenta e grande).

La memoria cache ha due tipi di applicazioni: la cache primaria, che è direttamente integrata nel processore, e la cache secondaria, la quale può essere esterna ad essa. Sono memorie di tipo volatili, dalle prestazioni elevate e di dimensioni ridotte. Per avere queste caratteristiche si cerca

di diminuire il numero dei transistor in modo tale da far abbassare la grandezza delle celle è magari anche di far incrementare le prestazioni.

SRAM 5T

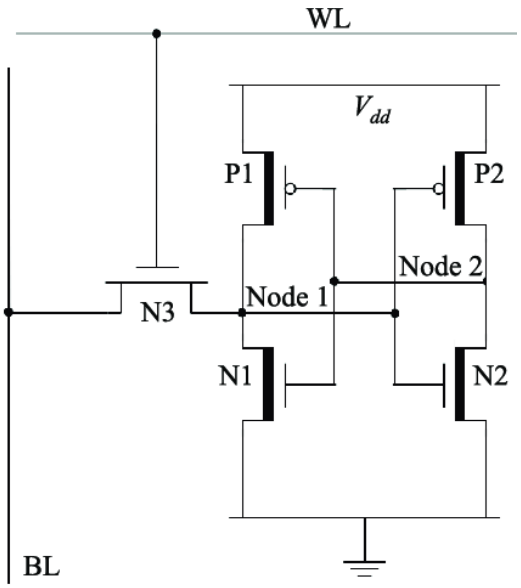


Figura 7: Cella SRAM 5T

Nella Figura 7 è illustrato il circuito di una cella SRAM a cinque transistors è una cella a terminazione singola che consiste in un solo pass transistor invece di due. È costituito da una sola bit line, quindi i dati possono essere scritti o letti solo attraverso essa. Avendo eliminato un transistor, le dimensioni della cella si riducono e le prestazioni potrebbero aumentare. Nella pratica questo tipo di configurazione non viene particolarmente utilizzata perché abbiamo bisogno di modificare l'architettura per poterla utilizzare.

Per quanto riguarda il funzionamento, è lo stesso della configurazione a sei transistor.

SRAM 2T

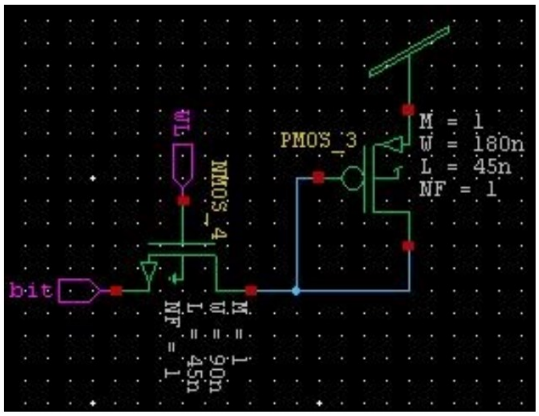


Figura 8: Cella SRAM 2T

Nella Figura 8 è illustrato il circuito di una cella SRAM con configurazione a due transistor. È composta da un NMOS e da un PMOS.

Essendo che il funzionamento di tale cella è molto diversa rispetto alla 6T, vale la pena illustrare il suo funzionamento e le sue problematiche:

- Scrittura di '1': durante la scrittura, quando è necessario scrivere un 1, NMOS trasferisce i dati da BL al gate e drain di PMOS. Il bit 1 non attiva mai il PMOS e i dati 1 devono essere memorizzati nella floating point tra il gate ed il drain. A causa della capacità parassita, il bit di dati 1 si degraderà nel tempo. Dopo alcuni nanosecondi, i dati potrebbero non essere più considerati logici 1 perché sono stati degradati. Pertanto, durante la lettura di questi dati, potrebbe non essere identificato come logico 1 (potrebbe verificarsi un errore di lettura)
- Scrittura di '0': nel caso del dato '0', esso accenderà il transistor PMOS, facendo sì che V_{DD} passi attraverso il PMOS e raggiunga il gate ed il drain. Questo V_{DD} può spegnere il transistor PMOS, dopo qualche fluttuazione in questo momento, i dati stabili potrebbero essere diversi da 0V. Pertanto, anche in questo caso abbiamo una degradazione del bit, portando ad un conseguente errore di lettura.

SRAM 3T

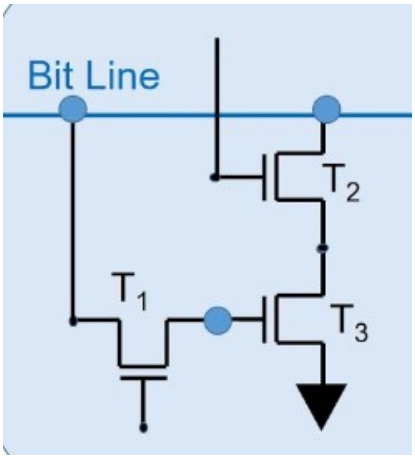
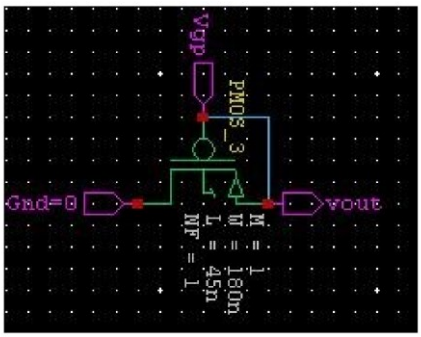


Figura 9: SRAM 3T

Questa configurazione, come possiamo osservare nella Figura 9, è composta da due transistor NMOS e da un transistor PMOS. La particolarità di questa cella è che usiamo un NMOS come pull-up ed un PMOS come pull-down.

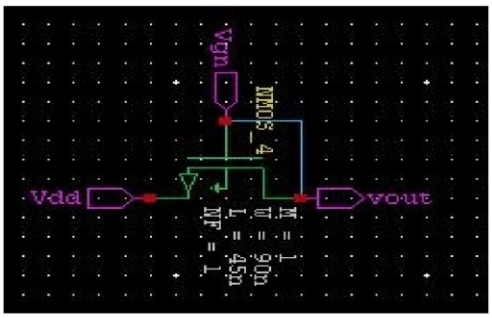
Il comportamento dei transistors in condizione di '1' o '0' è illustrato in Figura 10:

- Figura 10.a (Vgp=0V): Vout=GND -Vt, allora Vgp= GND -Vt. Questa tensione è sufficiente ad accendere il PMOS e Vout=-Vt che è un valore vicino a GND. In caso il valore logico è '0'.
- Figura 10.b (Vgp=V_{DD}): Vout= V_{DD}-Vt, allora Vgp= V_{DD} -Vt. Questa tensione è sufficiente ad accendere il NMOS e Vout=V_{DD}-Vt che è un valore vicino a V_{DD}. In caso il valore logico è '1'.



PMOS pass transistor with gate and source short circuited.

Figura 10.a



NMOS pass transistor with gate and source short circuited

Figura 10.b

Adesso andiamo ad analizzare il funzionamento di tale cella:

- Scrittura: durante l'operazione di scrittura WL=1 e BL rappresenta il valore che si vuole passare attraverso il pass transistor NMOS.
  - BL=0: il PMOS si accende e scarica l'output a GND. Ricordando che l'uscita è collegata al gate del transistor, il PMOS resta acceso continuamente ed il dato resta registrato nella cella.
  - BL=1: l'NMOS si accende e carica l'output a V_{DD}. Ricordando che l'uscita è collegata al gate del transistor, l'NMOS anche resta acceso continuamente (come visto per BL=0) ed il dato resta registrato nella cella.
- Lettura: quando WL=1, il dato immagazzinato nella cella passa a BL.

SIMULAZIONE E CONFRONTO

Tutti gli schemi sono stati realizzato tramite S-EDIT e verificati attraverso H-SPICE e T-SPICE in tecnologia a 45nm [15].

Design name	**N° of transistors**	**Speed of operation**	**Avg. Power consumption**
5T SRAM Cell	5	11.1 GHz	0.74 µW
2T SRAM Cell	2	8.0 GHz	3.86 µW
3T SRAM Cell	3	20.8 GHz	1.91 µW

Tabella 2

Dalla Tabella 2 riusciamo ad osservare che la configurazione 2T è impraticabile. Oltre al fatto che un alto consumo energetico ed una bassa frequenza di operatività, degrada il dato memorizzato, portando ad errori nel funzionamento della memoria.

Per quanto riguarda le altre due configurazioni, notiamo sia dei miglioramenti che dei peggioramenti. Rimuovendo due transistor abbiamo sicuramente una riduzione dell'area delle singole celle ed un aumento della velocità del 87.4%, ma allo stesso tempo abbiamo un aumento del consumo del 158.1%.

SRAM 8T

Si propone una configurazione esente dai disturbi durante la lettura, costituita da una porta per la scrittura ed una per la lettura. La cella è composta da otto transistor ed utilizza il rilevamento differenziale. Il requisito di progettazione conflittuale delle operazioni di lettura/scrittura in una cella SRAM 6T convenzionale viene eliminato utilizzando transistor di accesso di lettura/scrittura separati. In più, utilizzano transistor di dimensioni minime per ottenere un'elevata densità, sono soggetti a variazioni significative della tensione di soglia a causa del posizionamento casuale del drogante e della rugosità del bordo del gate.

FUNZIONE E PRESTAZIONI

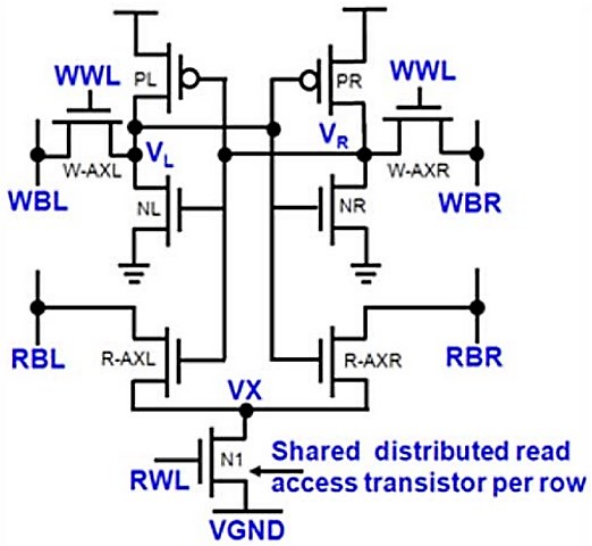


Figura 11: Cella SRAM 8T

La cella è mostrata in Figura 11. Osserviamo che, per ottenere il rilevamento differenziale con una eliminazione del disturbo in lettura, abbiamo bisogno di 8 transistor per cella e di un read-access transistor (N1) condiviso da tutte le celle della riga.

La stabilità di lettura nella cella 8T proposta è la stessa della stabilità in modalità hold. La porta di scrittura separata nella cella 8T proposta consente di aumentare le larghezze dei transistor di accesso in scrittura (W-AXL, W-AXR in Figura 11) rispetto al PMOS pull-up, migliorando così la scrivibilità.

In una cella da 6T convenzionale, per migliorare la stabilità di lettura, la larghezza dell'NMOS nella rete di pull-down è generalmente maggiore della larghezza del transistor di accesso, formando una diffusion notch. Le loro variazioni possono influire sulla resa, sulla dimensione della cella di bit, sulla densità di bit e sull'array. Se esse sono uniformi, si garantisce una migliore producibilità, un must per le future tecnologie su scala nanometrica.

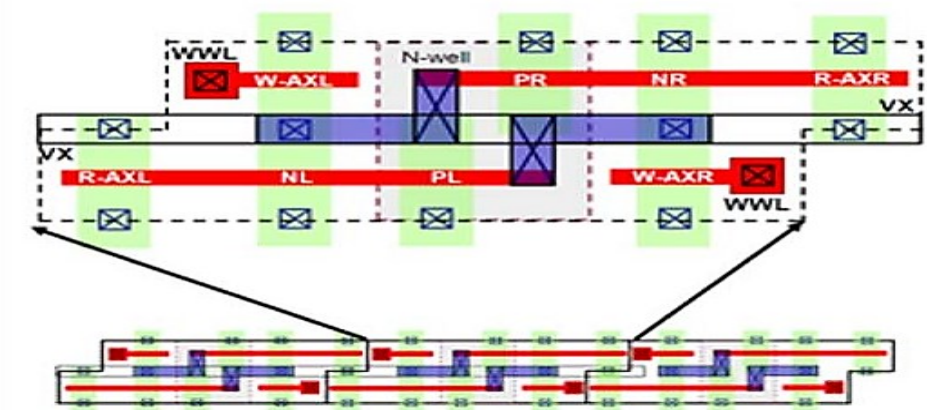


Figura 12: layout di celle sottili senza tacche di diffusione per la cella 8T

La Figura 12 mostra il layout di celle sottili senza tacche di diffusione per la cella 8T proposta insieme alle celle adiacenti. Ciascuna delle celle da 8T proposte richiede sei diffusion-notch, una delle quali è condivisa con la cella adiacente. Pertanto, la cella da 8T proposta richiede in media cinque tracce di diffusione (il layout convenzionale a cella sottile da 6T richiede quattro tracce di diffusione). La cella da 8T proposta ha un'area maggiore del 28% (compreso il transistor di accesso in lettura condiviso) rispetto alla cella da 6T di dimensioni minime.
Il transistor di accesso in lettura condiviso può ridurre la corrente di lettura a causa della capacità aggiuntiva sul nodo intermedio condiviso). Tuttavia, questo sovraccarico in termini di tempo di accesso può essere mitigato aumentando le dimensioni del transistor di accesso in lettura (N1).
Poiché il transistor di accesso in lettura (N1) è condiviso tra le celle della stessa riga, la perdita della cella di lettura (durante la modalità standby) può essere ridotta aumentando la tensione della sorgente (nodo VGND in Figura 11). Aumentando il potenziale VGND, la corrente di dispersione attraverso il transistor di accesso in lettura (N1) viene ridotta a causa dell'aumento della tensione di soglia (effetto body) e della tensione gate-source negativa. Si noti che il nodo VGND è diverso dal nodo VSS della coppia di inverter incrociati, e quindi l'alterazione della tensione VGND non influirà sulla stabilità della modalità di attesa della cella

CONFRONTO E CONCLUSIONI

Un chip di prova contenente celle 6T iso-area da 2 kb e gli array di cella 8T sono state fabbricate utilizzando la tecnologia CMOS da 90 nm [7]. L'array SRAM era organizzato con 256 righe e 8 colonne.

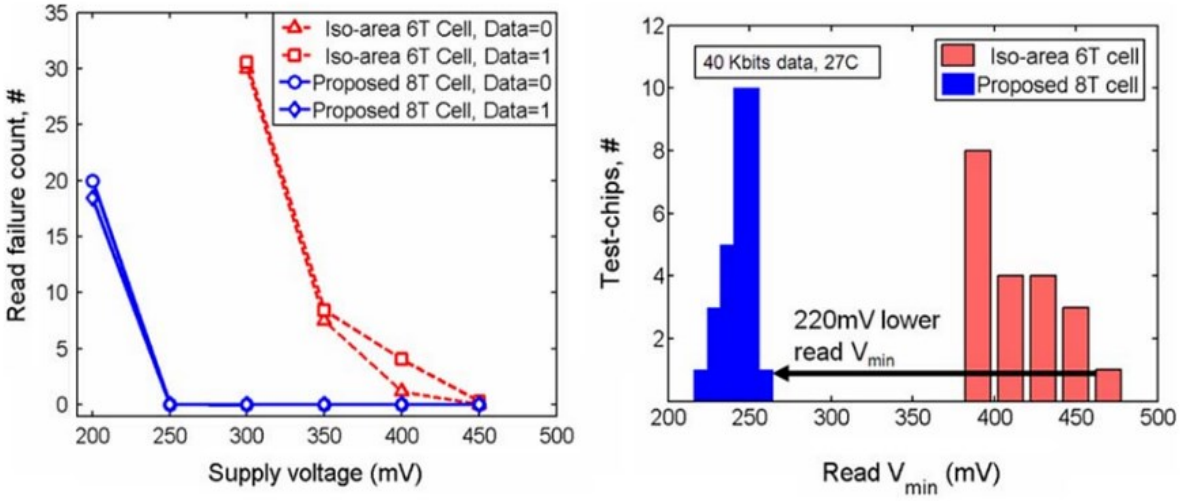


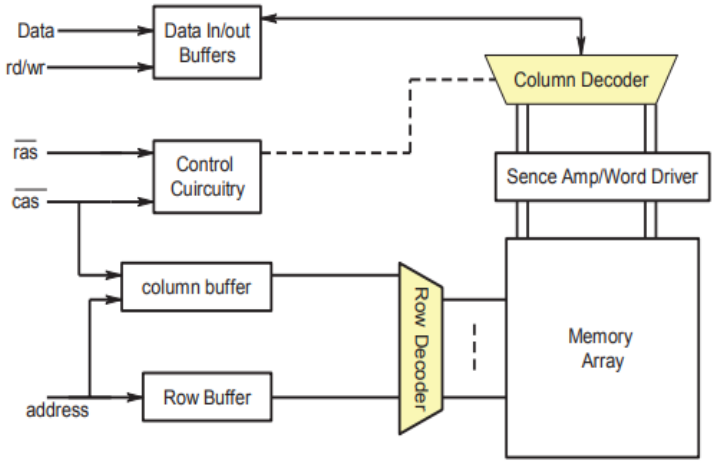
Figura 13: andamento della probabilità d'errore

La Figura 13 mostra gli errori di lettura di bit singolo misurati per '0' e '1' per la cella differenziale da 8T e la cella 6T iso-area calcolata in media su 20 misurazioni del chip di test. Per la 6T iso-area, sono stati osservati errori di mantenimento al di sotto di 300 mV. Quindi, per distinguere i fallimenti di lettura da quelli di attesa, le misurazioni dei fallimenti nella cella 6T sono state eseguite solo fino a 300 mV. Grazie al funzionamento privo di disturbi di lettura, la cella 8T mostra errori di lettura significativamente ridotti. Nella cella 8T, poiché la stabilità di lettura e la stabilità di attesa (stand-by) sono le stesse, gli errori di lettura inferiori a 250 mV sono essenzialmente errori di attesa.

La frequenza operativa è stata aumentata gradualmente (fino alla frequenza massima dell'analizzatore logico di 67,2 MHz) e sono stati osservati gli errori nel tempo di accesso. Poiché i transistor della porta di lettura nella cella 8T hanno le stesse dimensioni di quella della cella 6T iso-area, entrambe le celle hanno mostrato errori simili in termini di tempo di accesso. La corrente di dispersione è stata misurata mediante una misurazione diretta della sonda da una cella 6T iso-area da 32 kb e dagli array di celle 8T. La tensione del nodo VGND era controllata con un'alimentazione esterna. È stato osservato che aumentando la tensione del nodo VGND a 100 mV si riduce la corrente della bitline di lettura nella cella 8T del 70%. Inoltre, la cella 8T ha mostrato una perdita totale inferiore del 60% rispetto alla controparte 6T iso-area (a 600 mV, 125°C). La maggiore corrente di dispersione nella cella 6T iso-area è dovuta ai transistor pull-down sovradimensionati.

DYNAMIC RANDOM ACCESS MEMORY

Nelle SRAM considerate, l'unico scopo della rete di pull-up è quella di compensare le correnti di perdita dei transistor della cella. Una soluzione per eliminare il suo utilizzo è quella di compensare la carica persa con una riscrittura periodica del contenuto tramite un refresh. Questa operazione deve avvenire abbastanza frequentemente, in modo da non perdere l'informazione immagazzinata (tipicamente di 1-4ms). Il refresh è composto da due operazioni concatenate: abbiamo una prima fase di lettura del dato, con una conseguente scrittura.
Un funzionamento di questo tipo è definita Dinamico (DRAM), poiché il concetto alla base del suo funzionamento è basato sulla carica immagazzinata in una capacità.
Questo tipo di memoria ha un'ampia varietà di applicazioni, dai sistemi embedded, console per video games, workstation, supercomputers ecc... Le DRAM convenzionali utilizzano un meccanismo di divisione degli indirizzi, ma con un bus dedicato. Questa tecnica è usata nella maggior parte delle DRAM.



Architecture View of Conventional DRAM

Figura 14: Architettura di una DRAM

Come possiamo vedere nella Figura 14, nell'indirizzamento DRAM standard, il bus degli indirizzi è un multiplexer tra riga e colonna. Per indicare il core del chip DRAM quale (se uno dei due) di questi segnali viene pilotato sul bus degli indirizzi, vengono inseriti i segnali stroboscopici dell'indirizzo di riga e colonna, rispettivamente RAS e CAS, consentendo ai valori appropriati di essere fissati nel core.
Per leggere o scrivere una cella nella DRAM, il circuito di controllo calcola inizialmente il numero di riga del dato e lo invia ai pin del circuito di ingresso della memoria. Successivamente, attiva il pin RAS (Row Address Strobe), che ordina alla DRAM di leggere l'indirizzo di colonna presente in ingresso. La DRAM collega internamente la riga a una serie di amplificatori noti come sense amplifier, che leggono il contenuto di tutti i condensatori della riga e ne effettuano il refresh. Successivamente, il circuito di controllo invia il numero di colonna ai pin del circuito e attiva il pin CAS (Column Address Strobe), facendo leggere al circuito l'indirizzo di colonna del dato. La DRAM utilizza l'indirizzo di colonna per identificare il dato necessario dall'output

degli amplificatori. Dopo un intervallo di tempo chiamato CAS access time, il dato viene trasmesso all'esterno tramite il pin I/O della DRAM.
Dopo ogni operazione di lettura o scrittura, il circuito di controllo riporta i pin della memoria allo stato di riposo per prepararsi alla successiva operazione. La DRAM richiede un intervallo di riposo chiamato precharge interval tra un'operazione e la successiva. Questo intervallo di riposo è necessario per garantire che la DRAM si ripristini correttamente prima di eseguire una nuova operazione. Le DRAM multibit, che spesso hanno una larghezza di quattro o otto bit, consentono il funzionamento di più insiemi di celle contemporaneamente. Ogni insieme è dotato di un pin I/O e permette il trasferimento simultaneo di più bit di dati. Ciò consente un trasferimento dati più veloce rispetto alle DRAM a singolo bit. Tuttavia, anche le DRAM multibit richiedono ancora un intervallo di riposo tra le operazioni per il refresh dei dati.

DRAM 3T

La prima cella di memoria realizzata a funzionare in questa maniera è composta fa un circuito di tre transistor NMOS [16]. Questa configurazione tra largo utilizzo ai giorni d'oggi in molte memorie interne per specifiche applicazioni, anche se sono state ideate nuove celle con occupazione di area minori. Questo è dovuto alla sua semplicità.

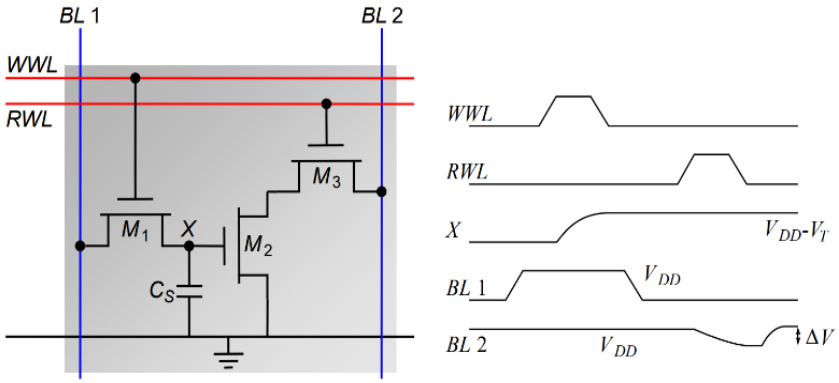


Figura 15: Cella DRAM 3T e andamento delle tensioni

In Figura 15 [18] possiamo vederla rappresentata: essa è costituita da tre transistor NMOS (M1, M2, M3) e di una capacità di memorizzazione (Cs), all'interno del quale sarà immagazzinato il dato. Bisogna precisare che la capacità non è un componente a parte, ma è la capacità di gate del dispositivo. Notiamo che non abbiamo più la presenza di una sola Word line, ma bensì due: una per la lettura (RWL) ed una per la scrittura (WWL). Le due Bit line, invece, vengono utilizzate per operazioni diverse: la prima viene utilizzata per la scrittura, mentre la seconda per la lettura.
Nella Figura 14 possiamo anche osservare il funzionamento della cella.

Durante la fase di scritto, imponiamo il valore desiderato alla Bit line 1 e, una volta che essa è stabile, abilitiamo la linea di scrittura WWL (il transistor M1 viene acceso). Da questo momento, la BL1 carica la capacità Cs della cella con un andamento esponenziale dettato dalla legge di carica/scarica del condensatore. In Figura 14 abbiamo la scrittura di un '1' logico, e possiamo notare che la capacità si carica fino a V_{DD}-Vt. Questo perché l'operazione viene effettuata tramite NMOS che '1' deboli. La caduta di tensione pari Vt riduce la corrente che passa attraverso il transistor M2, provocando un aumento del tempo di accesso. Per evitare ciò, la tensione V_{WW} applicata alla WWL è maggiore rispetto a V_{DD}.

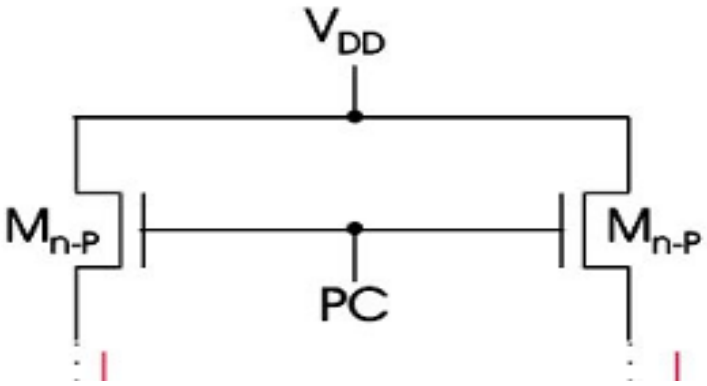


Figura 16: Dispositivo di precarica controllato da un segnale PC

Per la fase di lettura viene abilitata la linea RWL, la quale può attivare il transistor M3. Per quanto riguarda l'NMOS M2, la sua attivazione dipende dal valore immagazzinato della cella. La BL2 viene precaricata, attraverso un dispositivo apposito dispositivo illustrato nella Figura 16 (in questo caso controllato da un segnale PC), ad una tensione pari a V_{DD}-Vt. La lettura del valore immagazzinato viene effettuata sulla variazione di questa tensione.

Se la capacità Cs memorizza un '1', i transistor M2 e M3 scaricheranno la BL2 a massa, altrimenti resterà invariato. È evidente che la cella è invertente, cioè il valore letto è opposto a quello memorizzato al suo interno. Questo è un problema per la fase di refresh, in quanto bisognerà invertire il segnale letto prima di scriverlo nuovamente.

DRAM 1T

L'industria delle memorie DRAM (Dynamic Random Access Memory) ha compiuto miracoli rendendo sempre più piccole le celle di memoria. Sacrificando delle proprietà della cella DRAM a tre transistor, è possibile ridurre la complessità ed aumentare la densità della memoria.

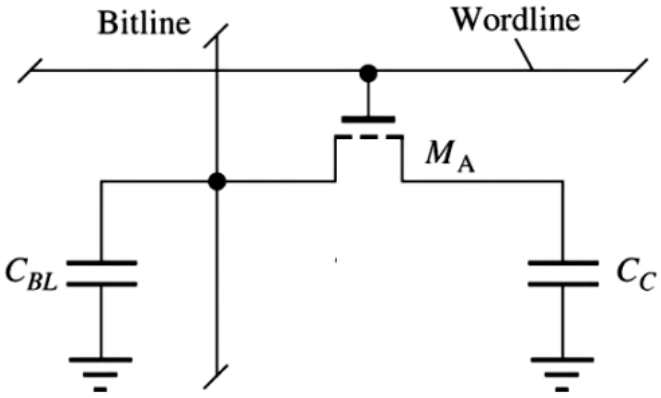


Figura 17: Cella DRAM 1T

La cella di memoria DRAM mostrata in figura 17 è composta da un singolo transistor MOSFET a canale n, chiamato transistor di accesso, e un condensatore di memorizzazione Cc [6]. Questa cella è nota come cella a singolo transistor, che la distingue dalle celle più vecchie che utilizzavano tre transistor. Il gate del transistor è collegato alla linea di controllo, Word line, mentre il source (o drain) è collegato alla Bit line. Il bit di informazione viene memorizzato come carica sul condensatore Cc. A causa delle perdite, il condensatore tende a scaricarsi nel tempo, quindi anche in questo caso è necessario un refresh periodico. Durante il refresh, il contenuto della cella viene letto e il bit letto viene riscritto, ripristinando così la tensione corretta del condensatore. Di solito, l'operazione di refresh viene eseguita ogni 5-10 ms.

Esaminiamo l'operazione di lettura.

Il primo step è quello di precaricare la Bit line ad un valore pari a $V_{DD}/2$. A questo punto portiamo a V_{DD} la Word line affinché il transistor M_A entri in conduzione.

Così facendo, il condensatore Cc ed il condensatore della BL C_B sono connessi in parallelo, come mostrato in Figura 18.

Figura 18: Capacità equivalenti della DRAM 1T

Valori tipici sono dell'ordine di: Cc 20-30 fF

C_B 200-300 fF

La rilevazione del valore immagazzinato viene effettuato sulla variazione della tensione ai capi di C_B. Nel caso sia immagazzinato un '1' (la tensione sul condensatore Cc è circa V_{DD})

avremmo un aumento della tensione di precarica. Per trovare il cambiamento in tensione sulla bit line risultante dal fatto di aver connesso il condensatore Cc ad esso, usiamo la legge di conservazione della carica elettrica:

$$C_C V_{CS} + C_B \frac{V_{DD}}{2} = (C_B + C_S)(\frac{V_{DD}}{2} + \Delta V)$$

tenendo conto nella seconda metà dell'equazione del parallelo tra le due capacità. Risolvendola otteniamo che

$$\Delta V = \frac{C_C}{C_B + C_C}\left(V_{CC} - \frac{V_{DD}}{2}\right)$$

tenendo conto che la capacità della BL è molto più grande di quella della cella Cc, possiamo riscrivere la formula

$$\Delta V \simeq \frac{C_C}{C_B}\left(V_{CC} - \frac{V_{DD}}{2}\right)$$

Tale variazione di tensione sarà rilevata da un sense amplifier, il quale si occuperà di amplificare il segnale caricando o scaricando la bit line. Questa operazione è di fondamentale importanza poiché la fase di lettura è un'operazione distruttiva e, se non si facesse quest'operazione di riscrittura, si perderebbe il dato. Contemporaneamente alla fase appena descritta, il segnale viene prelevato dalla linea dati in uscita del chip attraverso l'azione del decoder di colonna.

L'operazione di scrittura avviene in modo simile a quella di lettura. Si carica la Bit line alla tensione corrispondente al dato da scrivere. In seguito, viene abilitata la Word line, portandola a V_{DD}, affinché il transistor entri in conduzione. Ricordiamo che il condensatore Cc viene caricato tramite un NMOS; perciò, verrà portata ad una tensione pari a V_{DD}-Vt. Per risolvere questo problema possiamo aumentare la tensione applicata al gate dell'NMOS (pari a V_{DD}+Vt). Anche se il refresh viene effettuato automaticamente attraverso le operazioni di scritture e lettura, è comunque necessario effettuare il refresh completo dell'interno chip. Quest'operazione viene effettuata in una modalità a scoppio, una riga alla volta. Durante questo intervallo di tempo la memoria non è operativa, ma questo non è molto importante in quanto consiste solo nel 2% del tempo totale.

MEMORIA DRAM 3T1D

A causa del condensatore nella cella DRAM 1T che è diventato più difficile da ridimensionare, poiché le geometrie dei dispositivi si restringono. Recentemente le DRAM a tre transistor e un diodo (3T1D) senza condensatore (diodo con gate) hanno attirato l'attenzione, grazie alla loro capacità di raggiungere una maggiore densità e di risolvere i problemi associati al ridimensionamento del condensatore.

Nella Figura 19 possiamo vedere rappresentata la cella della memoria DRAM 3T1D.
La cella è composta da tre transistor NMOS, T1, T2 e T3, disposti come nella DRAM 3T, e da

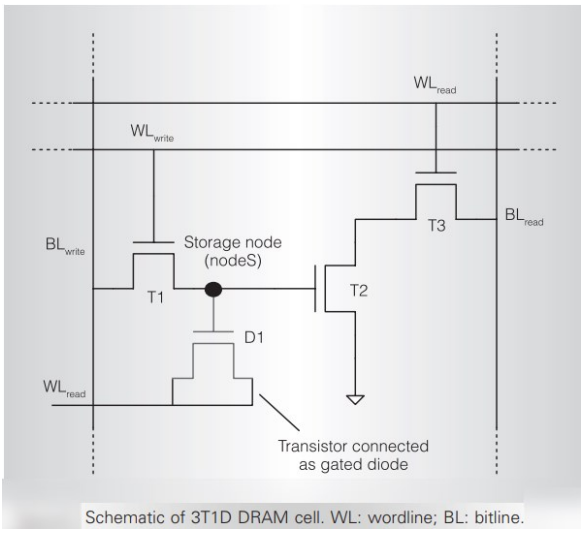


Figura 19: Cella DRAM 3T1D

un quarto transistor D1 in configurazione gated diode. Esso è un dispositivo a semiconduttore che combina la funzione di una giunzione p-n e di un condensatore MOS. È caratterizzato da un guadagno di amplificazione di potenza molto elevato. Considerando l'elevata impedenza di ingresso e la bassa impedenza di uscita, il diodo controllato dal gate agisce come una sorgente di tensione controllata dalla tensione. Come mostrato in figura, il gated diode si comporta come un noto capacito nel quale si memorizza il dato.
Portando ad un valore alto la WL write e tenendo bassa la WL read, si ha l'operazione di scrittura attraverso T1. Essendo che, anche in questo caso abbiamo la carica tra NMOS, dobbiamo adottare una tensione per WL write maggiore alla tensione di alimentazione (V_{DD}+Vt). Per l'operazione di lettura, come abbiamo visto per le altre configurazioni, abbiamo bisogno di precaricare la Bit line read ad un valore alto e portare la Word line read a V_{DD}.
Per la lettura di un '1', la capacità di D1 (Cgs) è maggiore del canale on, la tensione immagazzinata al nodo è alta ed accende il transistor T2. Questo porta a scaricare la BL read a ground attraverso la serie dei transistor T2-T3. Anche in questo caso, come nella configurazione 3T, la lettura è invertita: quando è immagazzinato un '1' la tensione sulla BL read diminuisce, mentre quando è presente uno '0' la BL rimane al suo valore di precarica.
I dati memorizzati in una cella DRAM basata su condensatore non possono essere conservati indefinitamente, poiché la corrente di dispersione alla fine rimuove o modifica i dati memorizzati. Lo scopo principale della variazione tecnologica è determinare l'efficienza, la dissipazione di potenza e la corrente di dispersione delle celle DRAM 3T-1D.
Quando la tensione applicata viene mantenuta costante a 4 V e la tecnologia cambia da 0,18 µm a 1 µm per la stessa architettura DRAM 3T-1D, otteniamo una corrente di dispersione diversa per tecnologia diversa, com'è possibile osservare dalla Figura 20.a.

Nella figura 20.b, invece, si mantiene costante la dimensione a 0.18µm e si fa variare la tensione da 1.8V a 5V. Si nota che all'aumentare della tensione, la dissipazione varia in maniera esponenziale [18].

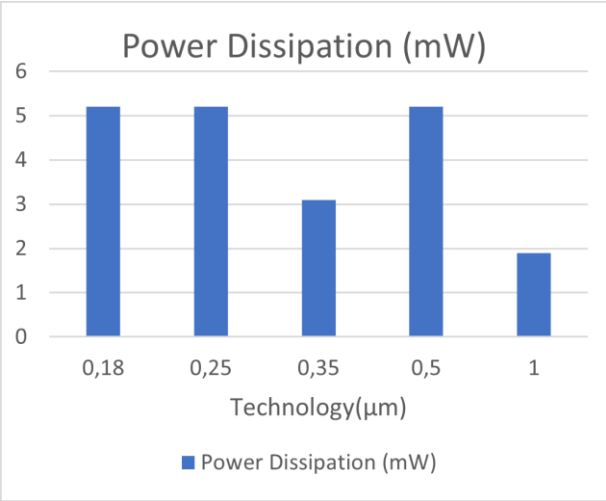


Figura 20.a

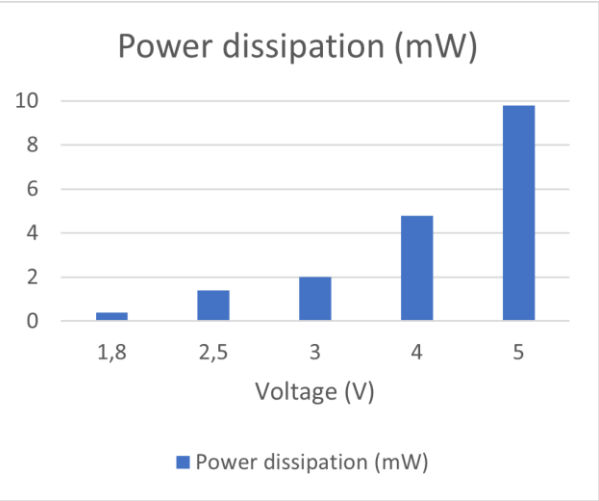


Figura 20.b

CONFRONTO TRA 6T SRAM E 3T1D DRAM

Considerando le memorie cache, gli squilibri del circuito dovuti alla mancata corrispondenza possono compromettere la stabilità delle celle e peggiorare le perdite nei tradizionali progetti SRAM. Inoltre, i numerosi percorsi critici e i pochi dispositivi in ciascun percorso contribuiscono ad aumentare la variabilità dei tempi di accesso. Quest'ultima caratteristica è di fondamentale importanza per i livelli più alti della gerarchia della memoria, che sono più critici per la latenza. La dipendenza della cella 6T dalla simmetria e dagli attenti requisiti di dimensionamento per adattarsi a letture e scritture robuste la rendono suscettibile a variazioni casuali. Una soluzione semplice a questi problemi è aumentare la dimensione della SRAM a scapito di prestazioni inferiori e area più ampia.
La configurazione 3T1D offre una valida alternativa alla memoria SRAM 6T [17].
Sebbene una cella 3T1D possa essere veloce, l'accesso ad alta velocità è valido solo per un periodo limitato dopo ogni scrittura sulla cella perché la carica si disperde nel tempo.

La Figura 22 traccia la relazione tra il tempo di accesso alla cella e il tempo trascorso dopo un'operazione di scrittura. Il tempo di accesso degrada fino a quando eccede il tempo di accesso della SRAM. Si definisce tempo di ritenzione come il periodo durante il quale la velocità di accesso può corrispondere con quello della 6T SRAM (circa 5.8µs), entro il quale è possibile accedere alla memoria alla frequenza nominale del chip.

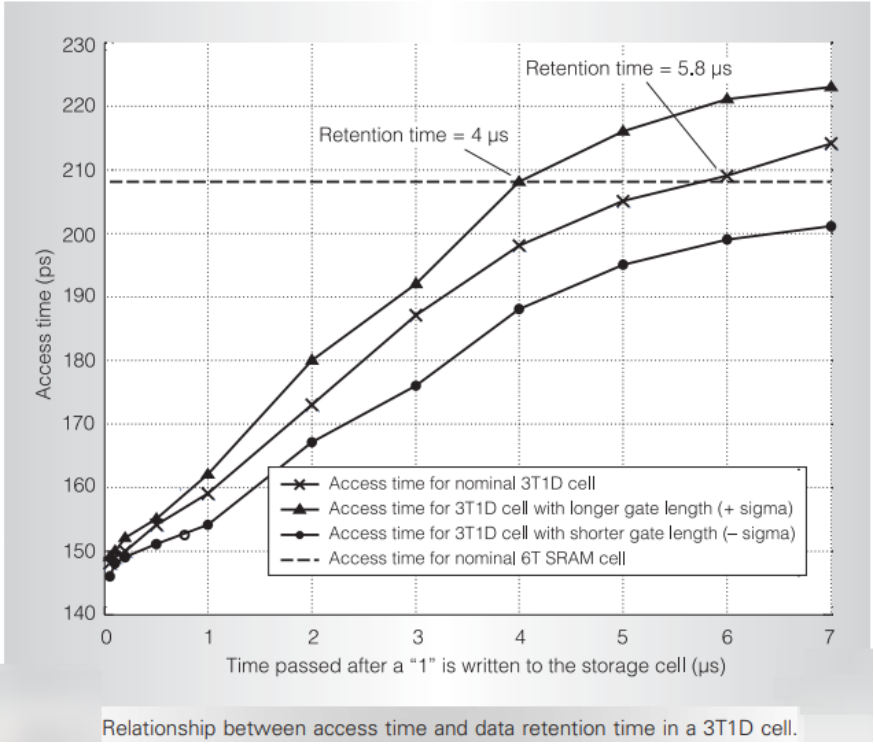


Figura 21: Grafico di confronto tra SRAM 6T e DRAM 3T1D

Una memoria 3T1D non è del tutto immune alla variazione del processo, ma un singolo parametro può assorbirne gli effetti e gestirli in modo efficiente. Ad esempio, se la variazione del processo riduce la capacità di pilotaggio del transistor di accesso T3, il tempo di accesso aumenta. Questo effetto può altrimenti essere visto come una diminuzione del tempo di ritenzione, come mostrato nella Figura 22. Il transistor di accesso più deboli hanno l'effetto di spostare la curva del tempo di accesso verso sinistra, riducendo il tempo di ritenzione della cella. Le variazioni del processo influiscono sulla frequenza operativa delle cache SRAM 6T. Al contrario, le variazioni del processo introducono variazioni del tempo di conservazione nelle cache 3T1D. Per ciascuna cache fabbricata, la cella di memoria con il tempo di conservazione più breve determina il tempo di conservazione dell'intera struttura.

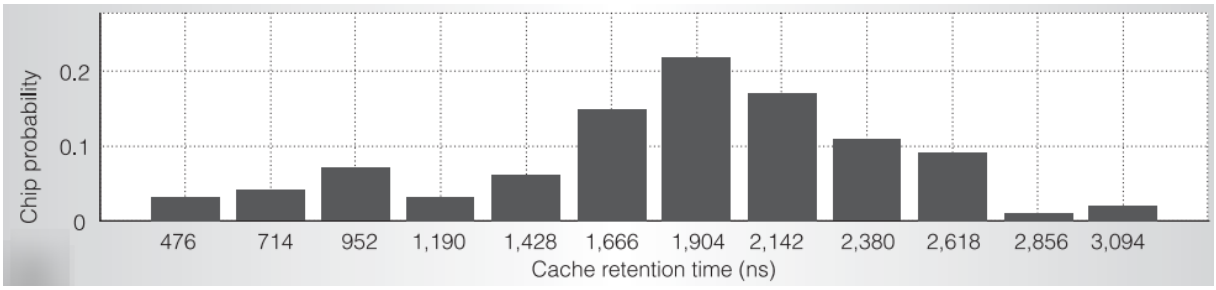


Figura 22

La Figura 22 presenta la distribuzione dei chip (o probabilità di produzione) come istogramma della distribuzione del tempo di ritenzione per la cache 3T1D con variazioni tipiche del processo. Sebbene tutti i chip siano progettati con gli stessi parametri, presentano un ampio intervallo di tempi di ritenzione a causa delle variazioni del processo.

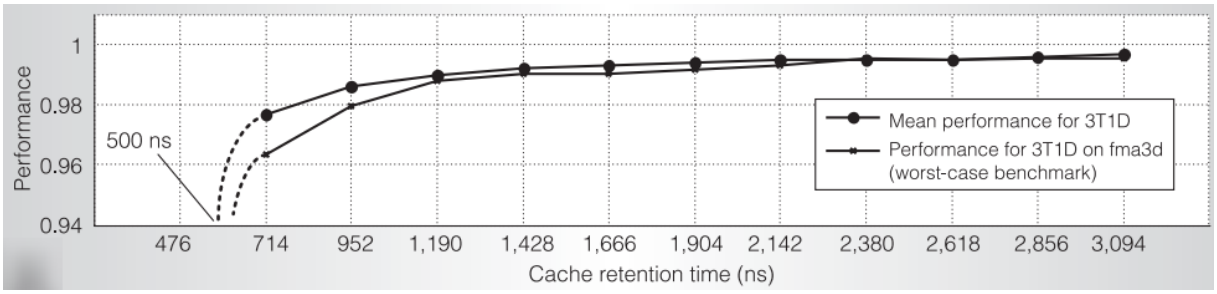


Figura 23

La Figura 23 mostra che le prestazioni del processore variano solo del 2% circa, con un tempo di ritenzione che varia da 714 ns a 3.094 ns. Il 97% delle cache 3T1D perde meno del 2% delle prestazioni rispetto a un design 6T ideale. Anche per il benchmark con le prestazioni peggiori, la penalità prestazionale è inferiore al 4%. Pertanto, una cache 3T1D raggiunge prestazioni migliori rispetto a un design SRAM con rendimenti comparabili. Tuttavia, questo aumento delle prestazioni va a scapito di un ulteriore consumo di energia dinamica per l'aggiornamento.

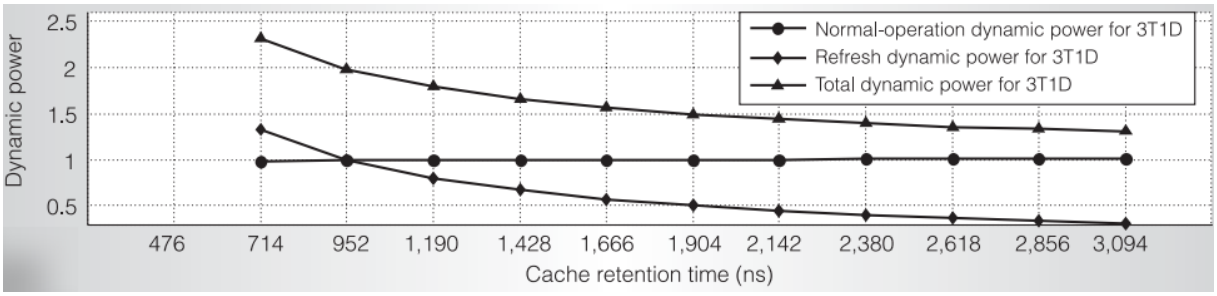


Figura 24

Il sovraccarico di potenza dinamica per il 3T1D varia da 1,3 a 2,25 volte quello del design ideale 6T, come mostra la Figura 24. Sebbene il consumo energetico dinamico sia maggiore, la potenza di dispersione tende a dominare le strutture della cache a 32 nm e, di conseguenza, si verifica una riduzione complessiva della potenza della cache rispetto al design SRAM.
Le celle 3T1D non soffrono dello stesso tipo di problemi di stabilità delle celle SRAM 6T, perché non vi è alcun conflitto intrinseco nelle celle 3T1D. Fatta eccezione per il tempo di conservazione dei dati limitato, una cella DRAM 3T1D è stabile. L'array di memoria basato su 3T1D offre anche vantaggi in termini di potenza di dispersione ridotta perché una cella 3T1D non soffre della moltitudine di forti percorsi di dispersione tra V_{DD} e terra riscontrati nelle celle SRAM 6T. Questo numero inferiore di percorsi di perdita porta a una perdita nominale inferiore e a una minore variabilità della perdita.

ANALISI PERFORMANCE TRA 3T E 3T1D

I parametri chiave per l'analisi delle DRAM sono: tempo di scrittura, tempo di lettura, dissipazione di potenza e periodo di ritenzione [18].

Tempo di scrittura

Per questo parametro consideriamo l'operazione di scrittura di un '1'.

Dalla Tabella 3 si nota che la DRAM 3T ha una migliore prestazione.

V_{DD} (tensione di alimentazione)	Write access time 3T DRAM in ps	Write access time 3T1D DRAM in ps
0.7	89.6475	236.4678
0.8	119.3147	285.4256
0.9	131.1353	332.36
1	146.7634	389.7869
1.1	177.6707	523.0148

Tabella 3

Questo è giustificator dal fatto che le capacità parassite di tale memoria sono inferiori rispetto alla corrispettiva cella 3T1D.

Tempo di lettura

Nella Tabella 4 si osserva che la configurazione 3T1D si comporta in modo migliore per ogni tensione.

V_{DD} (tensione di alimentazione)	Read access time 3T DRAM in ps	Read access time 3T1D DRAM in ps
0.7	----	590.6285
0.8	888.1687	138.1473
0.9	156.752	89.5566
1	95.3737	70.7436
1.1	74.6917	61.954

Tabella 4

Per una tensione di alimentazione pari o inferiore a 0.7V, la cella 3T non riesce ad eseguire l'operazione di lettura, mentre la cella 3T1D si comporta molto bene a tensioni basse (a 0.8V abbiamo una differenza di 750 ps tra le 2). Questo accade perché il gated diode aiuta l'operazione spingendo la tensione sullo storage node.

Dissipazione di potenza

V_{DD} (tensione di alimentazione)	Power dissipation 3T DRAM in µW	Power dissipation 3T1D DRAM in µW
0.7	0.193	0.4363
0.8	0.786	0.761
0.9	1.168	1.1940
1	1.696	1.7193
1.1	2.381045	2.4262

Tabella 5

Periodo di ritenzione

È il parametro più importante di una memoria DRAM. Esso rappresenta il periodo di tempo durante il quale il segnale immagazzinato deve essere valido per la lettura. Più grande è questo parametro e più piccola è la totale perdita di corrente della memoria. È evidente dalla Figura 6 che la 3T1D ha un periodo di ritenzione quasi doppio rispetto alla 3T.

V_{DD} (tensione di alimentazione)	Retention period 3T DRAM in µs	Retention period 3T1D DRAM in µs
0.7	16.5343	27.7
0.8	22.6875	38.1349
0.9	26.7371	42.7138
1	28.07	45.3586
1.1	28.23	48.7345

Tabella 6

Quando è memorizzato un '1', la perdita di corrente si traduce in una perdita di sottosoglia di T1 e in una perdita gate tunneling di T2. Per la richiesta di una maggiore densità e di migliori prestazioni delle memorie, la lunghezza del gate è ridimensionata continuamente, il che incrementa la perdita di corrente esponenzialmente.

In aggiunta, la perdita di sottosoglia di T1 è il fattore dominante per quanto riguarda il periodo di ritenzione.

$$I_{DS} = \mu_0 C_{ox} \frac{W}{L} (m-1)(v_T)^2 * e^{(V_g - V_{th})mv_T} * (1 - e^{\frac{v_{DS}}{v_T}})$$

dove

$$m = 1 + \frac{C_{dm}}{C_{ox}} = 1 + \frac{\frac{\varepsilon_{Si}}{W_{dm}}}{\frac{\varepsilon_{ox}}{t_{ox}}} = 1 + \frac{3t_{ox}}{W_{dm}}$$

V_{th} = tensione di soglia

v_T = tensione termica

C_{ox} = capacità dell'ossido del gate

μ_0=mobilità delle cariche

m = coefficiente effetto body

W_{dm} = larghezza massima dello strato di svuotamento

t_{ox} = spessore dello strato di ossido

C_{dm} = capacità massima dello strato di svuotamento

Per ridurre la perdita di sottosoglia, bisogna aumentare la lunghezza di T1 fino ad un certo valore, in modo da non compromettere l'operazione di scrittura, e i parametri di T2 e T3 variano per migliorare l'operazione di lettura.

EFFECT OF GATE LENGTH OF T_1 ON RETENTION PERIOD

Gate Length (nm)	L_{T1}=64	L_{T1}=50n	L_{T1}=42n	L_{T1}=32n
Retention Period (μs)	43.3259	33.0854	23.6884	5.0703

Figura 25

I risultati indicano che la bassa ritenzione in 3TID è principalmente dovuta alla perdita di sottosoglia (T1) dovuta a un transistor di accesso in lettura debole che fa sì che la maggior parte della carica fluisca nelle linee di bit. All'aumentare della lunghezza di T1 (transistor di accesso in scrittura), si verifica una sostanziale diminuzione del tempo di accesso in scrittura per aumentare il tempo di ritenzione, e con questa lunghezza di gate anche il tempo di accesso in scrittura della cella DRAM è accettabile.

T-RAM

Questa nuova tipologia di memoria RAM ha lo scopo di fondere l'elevata densità e la velocità delle memorie DRAM e SRAM. Questa tecnologia utilizza la resistenza differenziale negativa:

$$r_{diff} = \frac{dv}{di} < 0$$

La resistenza negativa è una proprietà di alcuni circuiti elettrici nei quali un aumento della tensione ai capi di un dispositivo dà luogo ad una diminuzione della corrente elettrica attraverso esso. La differenza rispetto ad una normale resistenza è la proporzionalità data dalla Legge di Ohm, la quale stabilisce un aumento della corrente all'aumentare della tensione. Una normale resistenza consuma potenza dalla corrente che la attraversa, mentre la resistenza negativa produce potenza.

THYRISTOR

Il thyristor, o Silicon Controlled Rectifier (SCR), è un componente simile ad un transistor, composto da multistrati di silicio. Il suo nome deriva dal materiale di cui è costituito, dal segnale di gate che lo controlla (controlled) e dal fatto che, una volta acceso, di comporta come un diodo raddrizzatore (rectifier).

In Figura 26 si può osservare il simbolo del componente e la sua struttura, la quale è composta da 4 strati di semiconduttore drogati P-N-P-N. Il Thyristor è un dispositivo unidirezionale che può funzionare come un interruttore a circuito aperto o come un diodo raddrizzatore a seconda di come viene attivato dal segnale di gate. In altre parole, si può usare il componente come un swith, ma non come un amplificatore.

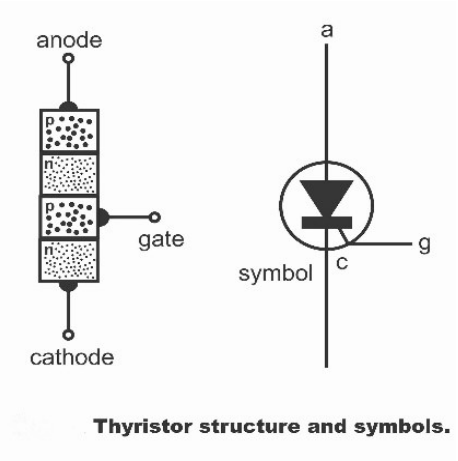


Figura 26: Thyristor

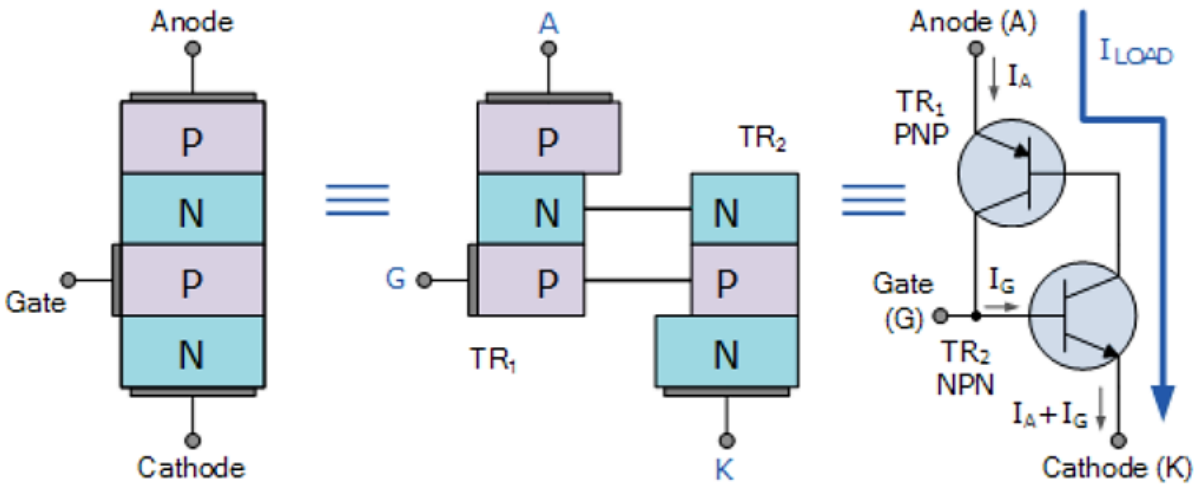


Figura 27: Thyristor come composizione di due transistor

Come illustrato nella Figura 27, è possibile considerare il Thyristor come una composizione di due transistor: TR1 e TR2. Questi due transistor interconnessi si affidano l'uno all'altro per la conduzione poiché ciascun transistor riceve la corrente di base-emettitore dalla corrente di emettitore-collettore dell'altro. Pertanto, finché a uno dei transistor non viene fornita una certa corrente di base, non può accadere nulla anche se è presente una tensione anodo-catodo. Quando la differenza di potenziale tra anodo e catodo è negativa, la giunzione NP centrale è polarizzata direttamente, ma le due giunzioni esterne PN sono invece polarizzate inversamente. In questo caso abbiamo il comportamento normale del diodo che, pertanto, blocca il flusso di corrente inversa. Se invece la differenza di potenziale è positiva, il componente blocca anche la corrente diretta.

Se viene iniettata una corrente positiva nella base del transistor TR2, la corrente di collettore risultante scorre nella base del transistor TR1. Ciò a sua volta fa sì che una corrente di collettore fluisca nel transistor TR1, che aumenta la corrente di base di TR2 e così via. Molto rapidamente i due transistor si costringono a vicenda a condurre fino alla saturazione poiché sono collegati in un circuito di feedback rigenerativo che non può fermarsi. Una volta attivato in conduzione, la corrente che scorre attraverso il dispositivo tra l'anodo e il catodo è limitata solo dalla resistenza del circuito esterno poiché la resistenza diretta del dispositivo durante la conduzione può essere molto bassa, inferiore a 1 Ω, quindi la caduta di tensione ai suoi capi, e anche la perdita di potenza, è bassa. Quindi possiamo vedere che un thyristor blocca la corrente in entrambe le direzioni di un'alimentazione AC nel suo stato "OFF" e può essere attivato "ON", funzionando come un normale diodo raddrizzatore, mediante l'applicazione di una corrente positiva al gate.

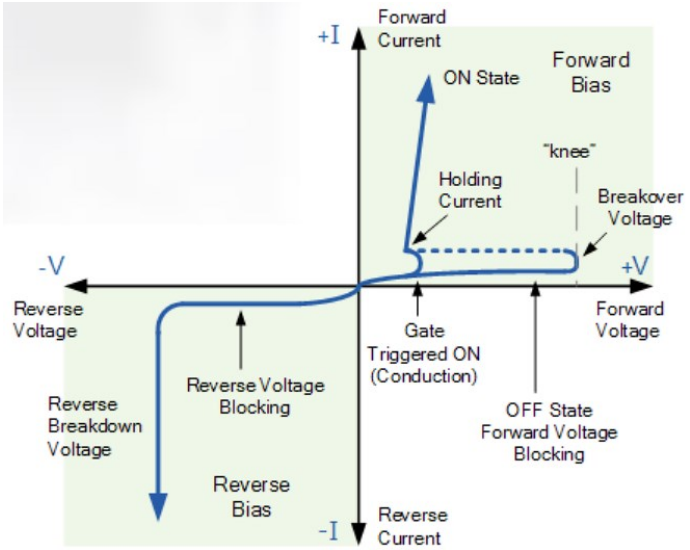


Figura 28: Caratteristica del thyristor

Una volta che il thyristor è 'ON' e fa passare corrente in avanti (anodo positivo), il segnale di gate perde tutto il controllo a causa del blocco rigenerativo dei due transistor. L'applicazione di eventuali segnali di controllo o impulsi dopo l'avvio della rigenerazione non avrà alcun effetto poiché il tiristore è già in conduzione ed è completamente attivo. Si può pensare il thyristor come un Latch bistabile avente due stati stabili: 'ON' e 'OFF'. In assenza di un segnale di gate,

il SCR blocca la corrente in entrambe le direzioni di una forma d'onda AC, mentre non può essere nuovamente 'OFF' usando il suo gate.
Per spegnere il componente bisogna rimuovere completamente la tensione di alimentazione, e quindi rimuovere o ridurre la corrente anodica mediante mezzi esterni al di sotto di un valore IH, chiamato "corrente di mantenimento minimo". Il mantenimento di questo deve avvenire abbastanza a lungo da consentire alle giunzioni PN agganciate internamente di recuperare il loro stato di blocco. Questo implica che, affinché un thyristor conduca, la sua corrente anodica IL deve essere maggiore del valore di corrente di mantenimento (IL>IH).

3-T TRAM

La riduzione delle dimensioni delle celle di memoria DRAM è in continua richiesta per maggiore densità, velocità e minor richiesta di potenza per la sua operatività. La tipologia di DRAM 1T-1C, come si è analizzato in precedenza, rende difficile diminuire la dimensione sotto una certa soglia per via della capacità. Per superare queste problematiche, è possibile utilizzare il thyristor per creare una memoria ad accesso casuale (TRAM).

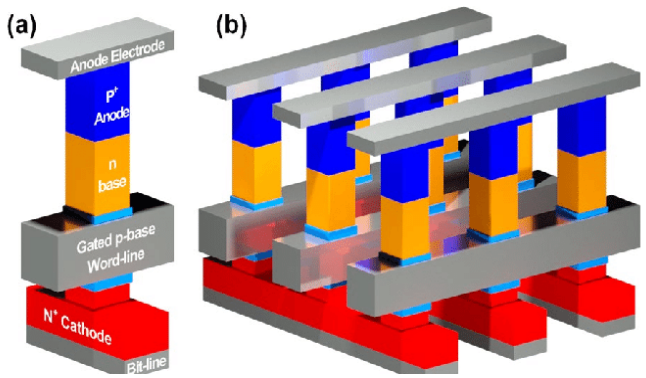


Figura 29: Cella e conessioni TRAM 3-T

Come illustrato nella Figura 29, la composizione della cella è costituita da P^+-N-P-N^+. La notazione con il + indica una maggior concentrazione di drogaggio rispetto alle zone senza (differiscono di circo di due ordini di grandezza). Questa simmetria nel drogaggio garantisce una caratteristica di isteresi della memoria. Ogni catodo della cella è collegato all'altro tramite la Bit line. Nello stato di standby, la tensione $V_{GC,ST}$ (tensione tra gate e collettore in standby) è tenuta ad un valore di -0.4V. La $V_{AC,P}$, invece, è tensione tra anodo e catodo per eseguire l'operazione di programmazione, la quale è pari a 1.2V con un V_{GC} pari a 0.4V. Questo valore di $V_{AC,P}$ è dovuto all'accumulo di portatori nella base P dalla tensione negativa di $V_{GC,ST}$.
Per l'operazione di scrittura, la tensione V_{GC} sale velocemente a 0.4V, facendo variare le lacune in P. Questo porta a ridurre l'altezza della barriera di potenziale nella base P (H_P).

In Figura 30 abbiamo la caratteristica tra corrente di anodo I_A e tensione V_{AC} con V_{GC} fissa a -0.4V [3]. Il dispositivo presenta un rapido incremento della corrente ad una tensione V_{AC} pari a 2.65V, il quale rappresenta il passaggio dallo stato '0' allo stato '1'. Questo indica l'esigenza di una tensione elevata se si mantiene V_{GC} ad una tensione così bassa. Se invece incrementiamo V_G a 0.4V, basta una tensione di 1.2V affinché si possa far commutare lo stato. In questo modo possiamo evitare errori di programmazione indesiderata, in quanto, tenendo V_G ad una tensione inferiore a -0.4V, è presente un range di 1.45V tra le tensioni di programmazione, il che rende queste celle molto robuste.

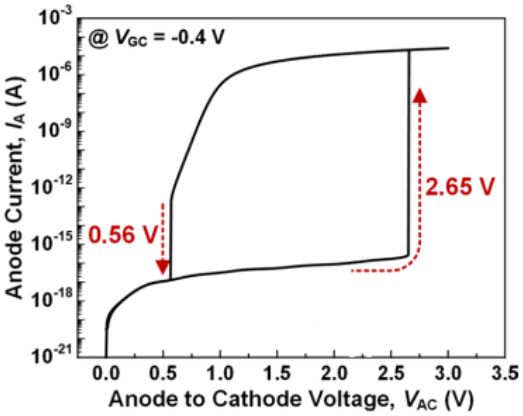


Figura 30: Caratteristica tra corrente di anodo I_A e tensione V_{AC} (V_{GC}= -0.4V)

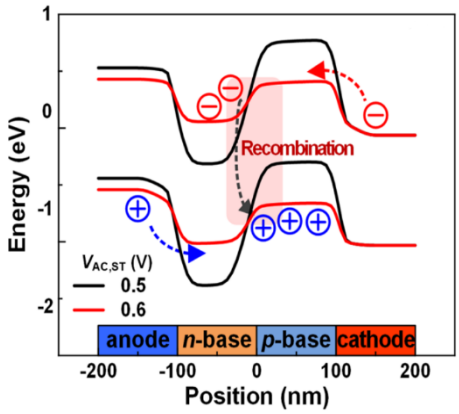


Figura 31: Andamento delle barriere di potenziale al variare di $V_{AC,ST}$

Si vuole analizzare adesso la variazione intorno a 0.56V. Per una tensione $V_{AC,ST}$=0.6V, nella regione di base vengono iniettate lacune in numero maggiore rispetto alla quantità ricombinata e vengono mantenute la forma della banda lo stato '1' insieme alla carica immagazzinata. Il dispositivo riesce a mantenere un numero di lacune elevato (10^{18} cm^{-3}) per più di 10 secondi.

Per una tensione $V_{AC,ST}$=0.5V, le lacune iniettate non sono sufficienti a compensare quelle che si ricombinano. Le lacune immagazzinate spariscono dopo 10µs ed il dispositivo torna allo stato '0'.

Per questo funzionamento, la cella non ha bisogno di refresh e, con una tensione $V_{AC,ST}$ di 0.6V, il dispositivo una corrente di standby pari a 1.14 pA, la quale la rende ideale per le operazioni a bassa potenza.

OPERATIVITÀ DELLA MEMORIA

Per una regolazione efficiente di V_{GC} nel funzionamento della memoria, gli elettrodi di gate e di catodo sono impostati rispettivamente su WL e BL e l'anodo mantenuto ad una tensione costante di 0.6V (Figura 32).

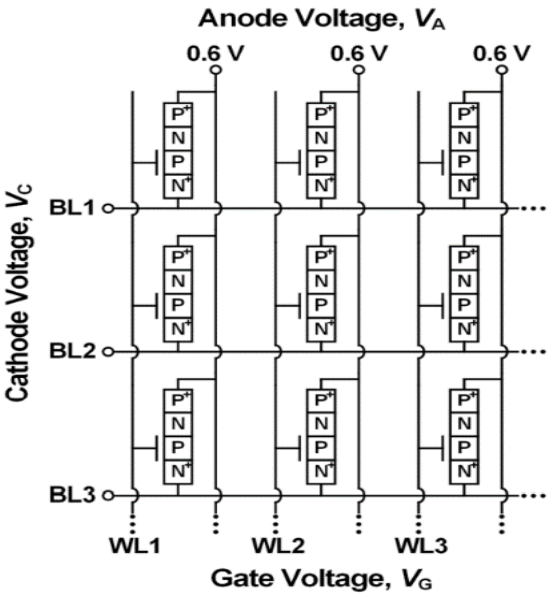


Figura 32: Matrice di celle TRAM 3-T

Si considerano le seguenti operazioni

- programmazione: per programmare una cella selezionata, la tensione V_C si abbassa fino a valore di -0.8V, portando V_{GC} a 0.4V (come visto in precedenza) e V_{AC} a 1.4V (maggiore di $V_{AC,ST}$ pari a 1.2V). Questo abbassamento facilita il flusso di lacune nella regione di base. Affinché non avvengano variazioni non desiderate nelle celle non selezionate, si portano le tensioni nelle WL non di interesse a -1.2V, così da avere una V_{GC} di -0.4V (portando la tensione $V_{AC,ST}$ a 2.65V). Alla fine, si avrà la cella selezionata nello stato '1', mentre le altre rimarranno immutate.
- Cancellazione: per la selezione della cella V_G passa da -0.4V a 0.8V, mentre V_C passa da 0V a 0.4V. Il numero di lacune nella regione P diminuisce per via dell'incremento di V_{GC}. In aggiunta, le lacune iniettate nella regione durante l'operazione di cancellazione sono bloccate dalla diminuzione della tensione V_{AC} che passa da 0.6V a 0.2V. Come per l'operazione di programmazione, anche qui si potrebbe riscontrare la possibilità di cancellare tutte le celle della WL. Per evitare questo comportamento indesiderato, dobbiamo ridurre la tensione V_{GC} a -0.4V portando V_C a 1.2V per tutte le BL non selezionate.

- Lettura parallela: la tensione della BL selezionata V_C e di tutte le WL V_G vengono portate a -0.8V. Per studiare l'effetto dell'operazione di lettura sullo stato ‘0’, la tensione operativa e la corrente anodica vengono estratte dopo l'applicazione di dieci operazioni di lettura consecutive a seguito di un'operazione di cancellazione come mostrato in Figura 33. Sebbene i dieci impulsi di lettura continui vengano applicati al dispositivo dopo l'operazione di cancellazione, la corrente di lettura diminuisce gradualmente, confermando che lo stato ‘0’ viene mantenuto stabilmente. Questo risultato indica che la 3-T TRAM con la configurazione di array suggerita per la lettura mostra un'affidabile immunità ai disturbi per lo stato ‘0’.

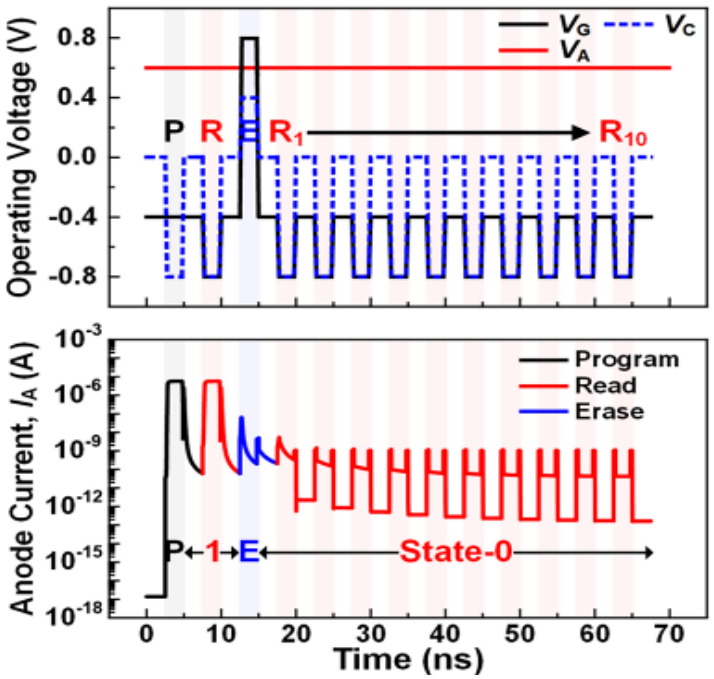


Figura 33: andamento della tensione e della corrente durante le operazioni di lettura, scrittura e cancellazione nel tempo

FLASH

Introdotta già a partire dal 1984, questa tipologia di memoria non volatile è una combinazione delle strutture EPROM (la sua densità) e EEPROM (la sua versatilità), infatti utilizza anch'essa la tecnologia a gate flottante [12]. Attraverso questo particolare dispositivo si riesce a far variare la tensione di soglia Vt immagazzinando cariche nel gate flottante attraverso l'effetto tunnel. Applicando infatti una tensione sufficientemente alta, le cariche hanno abbastanza energia da riuscire a saltare l'ossido di silicio per effetto tunnel, aumentando o diminuendo la tensione di soglia Vt. Le Flash sono praticamente delle memorie basate sui MOSFET che possono essere cancellate e riprogrammate elettricamente. Queste trovano utilizzo in tutte quelle applicazioni dove vogliamo che i dati vengo sia scritti che letti e dove si vogliono mantenere memorizzati anche dopo l'interruzione dell'alimentazione. Il suo primo impiego è stato nelle memory cards per trasferire ed immagazzinare fati tra computer ed altri dispositivi digitali. Bisogna comunque specificare che, nella maggior parte dei casi, vengono usate come memorie secondarie come hard disks.

Le principali tipologie di memorie Flash sono le NOR, le quali hanno una minor densità ma la possibilità di eseguire un accesso randomico, e le NAND, le quali hanno una maggiore densità.

Features	NOR	NAND
Memory size	≤512 Mbit	1 to 8 Gbit
Sector size	Approximately 1 Mbit	Approximately 1 Mbit
Program time	9 µs/word	400 µs/page
Erase time	1 s/sector	1 ms/sector
Read access time	<80 ns	20 µs
Write parallelism	8 to 16 words	2 Kbyte
Output parallelism	Byte/word/dword	Byte/word
Read parallelism	8 to 16 words	2 Kbyte
Access method	Random	Sequential
Price	High	Very low
Reliability	Standard	Low

Tabella 7

Nella NOR Flash, le celle sono collegate in parallelo alle linee di bit, il che permette in particolare di leggere e programmare le celle individualmente. Nella NAND Flash, invece, le celle sono collegate in serie, in modo simile a una porta NAND. Le connessioni in serie consumano meno spazio di quelle parallele, riducendo il costo della NAND Flash ed aumentandone la densità. Per quando riguarda i tempi delle prestazioni, la configurazione NOR richiede tempi di acceso più brevi per via della sua semplicità nella composizione. Questa sua particolarità la rende particolarmente utile sulle memorie embedded ed in tutte quelle applicazioni dove si richiede una maggior velocità in lettura. Essendo che possiede anche la capacità di esecuzione sul posto, è particolarmente usata per i microcontrollori. La configurazione NAND, invece, ha tempi di programmazione e di cancellazione minori. Questo la rende ideale per archiviare grandi quantità di dati, come per esempio hard disk, USB, memorizzazione di file video e audio ecc.…

Tuttavia, la NAND non può eseguire operazioni di lettura e scrittura contemporaneamente; può realizzarli a livello di sistema utilizzando un metodo chiamato shadowing, utilizzato da anni sui PC caricando il BIOS dalla ROM più lenta alla RAM ad alta velocità. La Tabella 7 evidenzia le principali differenze tra NOR e NAND [5]. Ciò dimostra che la NAND è ideale per l'archiviazione di dati ad alta capacità, mentre il NOR è utilizzato al meglio per l'archiviazione e l'esecuzione di codice, solitamente con capacità ridotte.

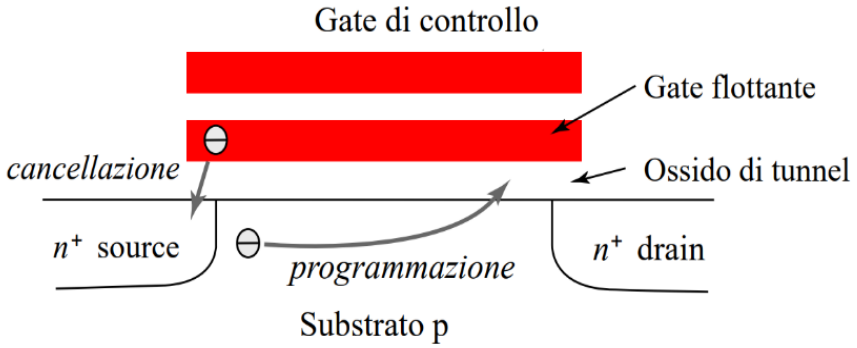


Figura 34: Cella Flash ETOX

La memoria Flash NOR opera nella seguente maniera, utilizzando delle Flash ETOX introdotta da Intel [19] come quella raffigurata in Figura 34. Questo dispositivo impiega un sottile strato di ossido di tunnel (10nm) come dielettrico di gate. Esso utilizza l'effetto Tunnel Fowler-Nordheim per scrivere dati nella cella. Questo principio si basa sulla probabilità non nulla per un elettrone di superare la barriera di potenziale sottoposto ad un campo elettrico. La probabilità di tunnel è legata alla distribuzione di cariche nel gate flottante, alla forma dell'ossido di tunnel e al suo spessore. Tutti questi aspetti compongono la "trasparenza di barriera". Questa presenza o assenza di elettroni nel gate flottante fa variare la tensione di soglia del dispositivo, distinguendo così i due valori logici '0' e '1'.

Per la cancellazione, si applica una tensione elevata al source (12V), tenendo a 0V il gate. In questo modo, gli elettroni presenti nel gate flottante sono iniettati verso il source, a potenziale maggiore, per effetto tunnel. Per evitare problemi sulla dispersione delle tensioni di soglia, dovuti a squilibri tra le celle, si programmano tutte le celle del blocco in modo che abbiano tutte la stessa tensione di soglia prima della cancellazione. In seguito, si applica una lettura di controllo per verificare che tutte le celle del blocco siano state effettivamente cancellate.

Per l'operazione di scrittura (o programmazione), si applica un impulso di tensione elevata (12V) al gate del dispositivo ed una determinata tensione (in base al dato che vogliamo immagazzinare) al drain. Se manteniamo una tensione 0V sul drain, non passa corrente attraverso il canale del transistor e, di conseguenza, non si possono iniettare elettroni nel gate flottante. Questa assenza di cariche corrisponde al valore logico '0' (dispositivo sempre acceso). Se invece al momento dell'impulso è presente una tensione al drain (6V), passa corrente sul canale, e gli elettroni accelerati riescono a passare nel gate, aumentando la tensione di soglia e, di conseguenza, trasformando il transistor in un dispositivo sempre spento. Questo stato equivale al valore logico '1'. Per ottenere la traslazione della tensione di sogli desiderata, è sufficiente un impulso di durata 1-10µs.

Per l'operazione di lettura basta portare la Word line corrispondente 5V, tenendo la Bit line a massa. In questo modo è permessa la selezione randomica.

Nella figura 35 è raffigurata la memoria Flash in configurazione NAND. Il modulo elementare consiste in una serie di 8-16 transistor a gate flottante connessi in serie. Ai capi della catena sono posti due transistor di selezione, uno connesso alla Bit line e l'altro alla Source line (massa).

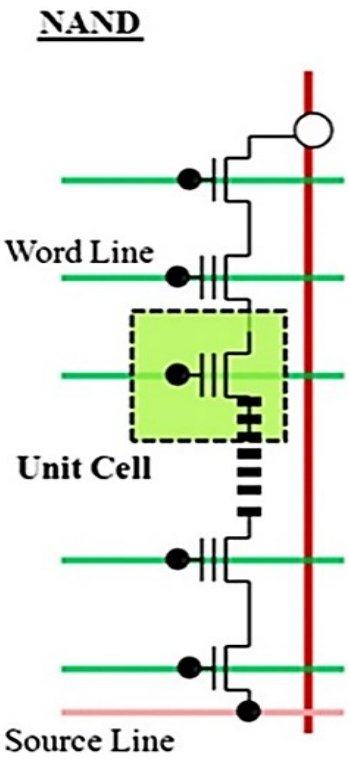


Figura 35: Memoria Flash configurazione NAND

L'operazione di programmazione ha inizio connettendo la catena di transistors alla Bit line, ma isolandolo dalla Source line, tramite i transistor di selezione. Per scrivere il valore logico '1', la BL viene connessa a massa, mentre si porta ad una tensione elevata (20V) la Word line di interesse. In questo modo entrano nel gate flottante gli elettroni, alzando la tensione di soglia Vt. Se invece vogliamo lasciare lo '0', portiamo la BL al valore elevato (20V), lasciando così la tensione di soglia inalterata. Durante questa fase non si ha flusso di corrente nel canale, il che comporta un consumo praticamente nullo di potenza statica. Tuttavia, servono elevate tensioni e tempi di programmazione. Bisogna perciò adottare delle tecniche ausiliarie, come l'utilizzo di pompe di carico, per ottenere tensioni elevate, e programmare più celle contemporaneamente, per minimizzare i tempi.

Per l'operazione di cancellazione, tutte le celle vengono programmate per lavorare in modalità di svuotamento, ossia con una tensione di soglia negativa. Questo avviene applicando una tensione elevata (20V) alla linea di bit e source, mentre alla Word line viene applicata una tensione nulla.

In fase di lettura, la Bit line di interesse viene precaricata ad una determinata tensione positiva, mentre la Source line viene portata a massa. La Word line di interesse viene portata alla tensione di lettura V_R, mentre le altre WL della BL sono portate una tensione tale da renderle accese indifferentemente dal loro stato (cancellati o programmati).

Se la V_R>Vt, il transistor di interesse è acceso, permettendo il passaggio di corrente tra BL e SL. Tramite un sistema di rilevazione di corrente, si assegna alla cella il valore logico '1'.

Se invece V_R<Vt, il transistor risulterà spento e non passerà corrente. Il sistema di rilevazione non misurerà nulla, e perciò assegnerà il valore logico '0'.

FeRAM è una RAM non volatile che combina l'accesso rapido in lettura e scrittura delle celle DRAM, costituita da una struttura di condensatore e transistor, come mostrato nella Figura 36.

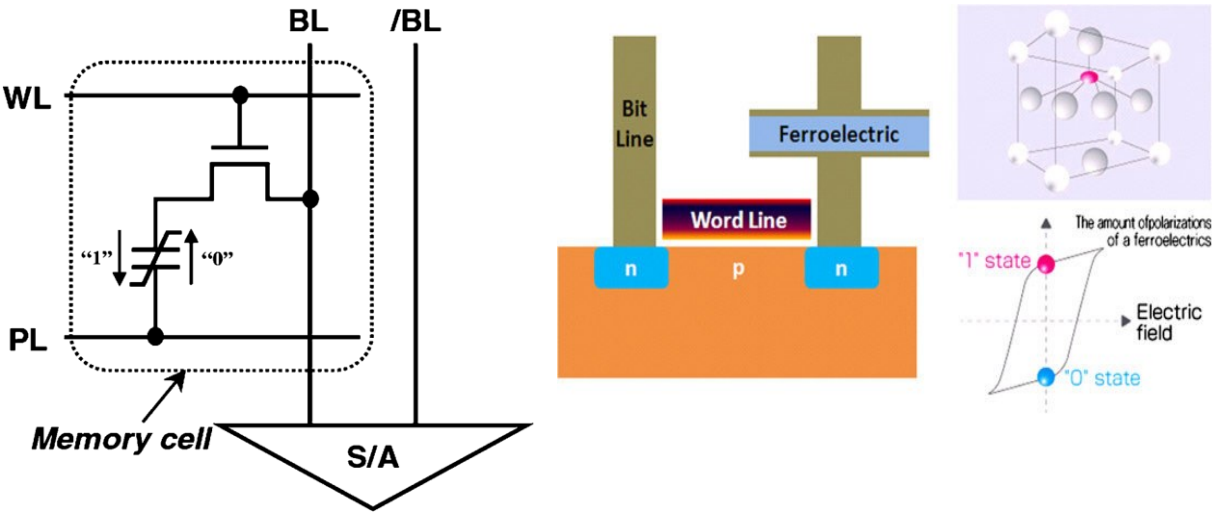


Figura 36: Memoria FeRAM (cella, componente ed insteresi)

Si accede quindi alla cella tramite il transistor, che consente di rilevare lo stato ferroelettrico del dielettrico del condensatore. Le proprietà di polarizzazione di una sostanza ferroelettrica vengono utilizzate come dispositivo di memoria. L'odierno FeRAM utilizza zirconato titanato di piombo (PZT) ma, ovviamente, si stanno valutando altri materiali in cerca di un miglioramento. FeRAM è il tipo più comune di memoria per personal computer con la capacità di conservare i dati quando l'alimentazione è spenta, così come altri dispositivi di memoria non volatile come ROM e memoria Flash. In una cella DRAM i dati necessitano periodicamente di un aggiornamento a causa dello scaricamento del condensatore, mentre FeRAM mantiene i dati senza alcuna alimentazione esterna. Ciò si ottiene utilizzando un materiale ferroelettrico al posto di un materiale dielettrico convenzionale tra le armature del condensatore. Quando un campo elettrico viene applicato attraverso materiali dielettrici o ferroelettrici, si polarizzerà e, mentre il campo viene rimosso, si depolarizzerà. Il materiale ferroelettrico, invece, mostra isteresi in un grafico di polarizzazione rispetto al campo elettrico e manterrà la sua polarizzazione. Uno suo svantaggio è che ha un ciclo di lettura distruttivo. Il metodo read prevede la scrittura di un bit su ciascuna cella; se lo stato della cella cambia, viene rilevato un piccolo impulso di corrente che indica che la cella era nello stato 'OFF'. Essa ha molte applicazioni in piccoli dispositivi di consumo come gli assistenti digitali personali (PDA), telefoni portatili, misuratori di potenza e smart card, nonché nei sistemi di sicurezza [5].

MRAM o RAM magnetica è una tecnologia RAM non volatile in fase di sviluppo dagli anni '90. I metodi RRAM per memorizzare i bit di dati utilizzano cariche magnetiche invece delle cariche elettriche utilizzate dalle tecnologie DRAM e SRAM. La MRAM si basa fondamentalmente su celle di memoria aventi due elementi di memorizzazione magnetici, uno con polarità magnetica fissa e l'altro con polarità commutabile. Questi elementi magnetici sono

posizionati uno sopra l'altro ma separati da una sottile barriera isolante a tunnel, come mostrato nella struttura cellulare in Figura 37.

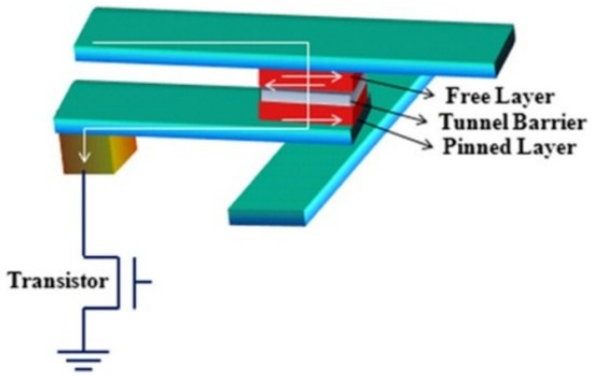


Figura 37: Architettura cella MRAM

Si definisce un metallo come magnetoresistivo se mostra un leggero cambiamento nella resistenza elettrica quando inserito in un campo magnetico. Combinando l'alta velocità della RAM statica e l'alta densità della DRAM, i sostenitori affermano che la MRAM potrebbe essere utilizzata per migliorare significativamente i prodotti elettronici memorizzando maggiori quantità di dati, consentendone un accesso più rapido e consumando meno energia della batteria rispetto alle memorie elettroniche esistenti. Tecnicamente funziona con lo stato della cella, che viene rilevato misurando la resistenza elettrica nel mentre una corrente attraversa la cella. A causa dell'effetto tunnel magnetico, se entrambi i momenti magnetici sono paralleli tra loro, gli elettroni saranno in grado di creare un "tunnel" e la cella si troverà nello stato "ON" a bassa resistenza. Tuttavia, se i momenti magnetici sono antiparalleli, la resistenza della cella sarà elevata.

Le caratteristiche di scrittura e cancellazione della memoria MRAM vengono soddisfatte facendo passare una corrente attraverso la linea di scrittura per indurre un campo magnetico attraverso la cella. MRAM è lentamente decollata, ma ora è entrata nel mercato e diventerà sempre più disponibile per la produzione di massa nel giro di un paio d'anni e oltre. Attualmente ha raggiunto un certo livello di successo commerciale in applicazioni di nicchia. Diverse aziende come Samsung, IBM, Hitachi, Toshiba e TSMC stanno sviluppando attivamente varianti di tecnologie dei chip MRAM. In termini di consumo energetico e velocità, le MRAM competono favorevolmente rispetto ad altre memorie esistenti come DRAM e Flash, con un tempo di accesso di pochi nanosecondi. Sebbene presenti alcune limitazioni durante l'operazione di "scrittura", la dimensione più piccola delle celle potrebbe essere limitata dalla diffusione del campo magnetico nelle celle vicine e necessitare di una riparazione per competere completamente come memoria universale.

MEMORIE RRAM

Attualmente un ulteriore aumento del numero di transistor non porta ad un aumento della velocità di clock o ad una riduzione del consumo energetico, ma solo ad una maggiore complessità e dimensione del circuito [4]. Per risolvere questo collo di bottiglia di von Neumann sono necessarie soluzioni tecnologiche innovative. Attualmente, due soluzioni principali vengono esplorate nei principali centri scientifici di tutto il mondo: combinare calcolo e memoria in un'unica unità funzionale e passare dalle tradizionali architetture di von Neumann alle architetture neuromorfiche, le quali riproducono i principi di archiviazione ed elaborazione delle informazioni nel sistema nervoso e nel cervello.

Il memristor (resistenza di memoria) è stato teoricamente descritto da Leon Chua nel 1971 [2] come un elemento passivo a due terminali mancante dei circuiti elettrici che mette in relazione la variazione del flusso magnetico φ(t) con la carica elettrica q(t).

$$q(t) \triangleq \int_{-\infty}^{t} i(\tau)d\tau$$

$$\varphi(t) \triangleq \int_{-\infty}^{t} v(\tau)d\tau$$

Questo componente può essere controllato da carica o dal flusso rispettivamente, dove φ(q) e q(φ) sono funzioni continue e differenziabili a tratti con pendenze limitate. Differenziando le due equazioni rispetto al tempo, si ottiene la resistenza (R) e la conduttanza (G) della memoria

$$R(q) \triangleq \frac{\mathrm{d}\varphi(\mathrm{q})}{dq}$$

$$G(\varphi) \triangleq \frac{dq(\varphi)}{d\varphi}$$

Tradizionalmente, i requisiti di archiviazione dati temporanei e permanenti di qualsiasi unità di elaborazione delle informazioni sono stati soddisfatti da memorie basate su MOSFET come DRAM, SRAM e memoria flash. La natura volatile della SRAM/DRAM porta alla perdita dei dati memorizzati non appena viene rimossa l'alimentazione. Di conseguenza, nel settore dei semiconduttori esiste una forte domanda di memorie non volatili (NVM) ad alte prestazioni, efficienti dal punto di vista energetico e ad alta densità.

Nel corso degli anni sono stati sviluppati diversi nuovi approcci a questa tipologia di memoria. Questi includono la memoria ad accesso casuale magnetica spin-transfer-torque (STT-MRAM), a cambiamento di fase (PCRAM), a ponte conduttivo (CBRAM) e resistiva (RRAM).

La RRAM offre molte caratteristiche vantaggiose come l'elevata velocità di commutazione, il basso consumo energetico e l'elevata densità. Inoltre, poiché la RRAM è compatibile con CMOS, CNFET [9] ed è altamente scalabile, è in grado di superare le prestazioni delle memorie convenzionali ad alte prestazioni come le SRAM convenzionali nel regime nanometrico. La resistiva, pertanto, sembra essere la tecnologia più promettente per le applicazioni di memoria.

Il memristor commuta tra uno o più valori di resistenze applicando livelli di tensione appropriati. Tali livelli distinti possono avere uno o più valori discreti oppure avere una

resistenza variabile continua, la quale è controllata dalla storia passata del dispositivo, ovvero dalla tensione precedente applicata al dispositivo. Il memristor è un resistore non lineare che cambia il suo stato di resistenza in base al flusso elettrico netto o alla carica netta che passa attraverso i suoi terminali, mantenendo il suo stato dopo aver rimosso la polarizzazione elettrica. Tale elemento di memoria di commutazione resistiva a base di ossido di metallo può essere utilizzato nelle RRAM grazie alla sua composizione semplice, al basso costo e alla compatibilità con la tecnologia CMOS. Inoltre, presenta anche un consumo energetico molto basso ed un'elevata densità.

La RRAM 1R [14] è la forma più comune di implementazione della memoria non volatile che utilizza memristor. Tuttavia, soffre di diversi problemi in termini di velocità, affidabilità e perdite dipendenti dai dati. Ad esempio, in una cella RRAM 1R, il problema della "corrente di dispersione" è un vincolo di progettazione importante che porta a una grande corrente di dispersione attraverso celle non selezionate nella riga e nella colonna selezionate e si traduce in un'operazione di lettura errata. Ciò avviene a causa dell'assenza di qualsiasi meccanismo di

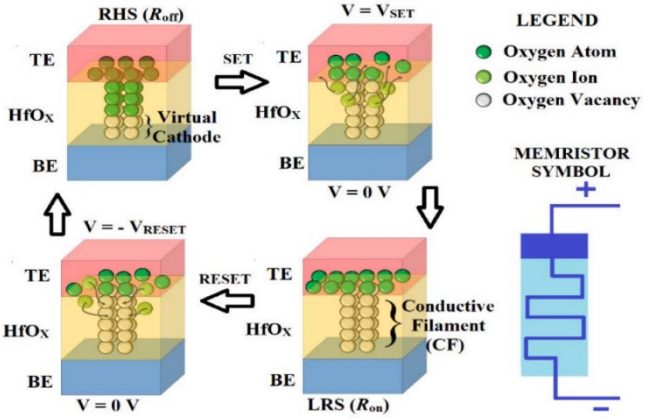


Figura 38: Funzionamento della cella RRAM

schermatura per isolare le celle non selezionate dalle tensioni di Bit line e Word line durante l'operazione di lettura.

Nel modello illustrato in Figura 38, HfOx è l'isolante, inserito tra un elettrodo superiore (TE) ed uno inferiore (BE), che commuta tra due stati resistivi: HRS e LRS, a seconda del la polarità del campo elettrico applicato ai suoi terminali. Si presuppone che il memristor memorizzi "1" quando si trova nello stato HRS (R_{off}) e memorizzi "0" quando si trova nello stato LRS (R_{on}). Durante il processo SET, viene applicata una tensione positiva ai terminali del memristor (memorizzando '1'), che porta alla deriva gli ioni di ossigeno attraverso l'anodo sino al HfOx. Ciò lascia dietro di sé un percorso conduttivo di ossigeno vacante attraverso l'ossido e, di conseguenza, il memristor viene portato nello stato LRS. D'altra parte, durante il processo di RESET, viene applicato un campo elettrico negativo ai terminali del memristor (memorizzazione di "0"). Ciò spinge gli ioni di ossigeno immagazzinati nel serbatoio verso il bulk HfOx che si traduce nella loro ricombinazione, porta il memristor nello stato HRS.

3T2R RRAM

La cella RRAM 3T2R [14], illustrata in Figura 39, è costituita da due memristor, MEM1 e MEM2, di polarità opposta, posizionati una di fronte all'altra. MEM1 e MEM2 sono collegati alla linea dati (DL) tramite pCNFET1. nCNFET1 e nCNFET2 collegano MEM1 e MEM2 alle corrispettive Bit line (BL e BLB). La Word line (WL) controlla i transistori di accesso nCNFET1 e nCNFET2 mentre il pCNFET1 è attivato dal WLB su riga.

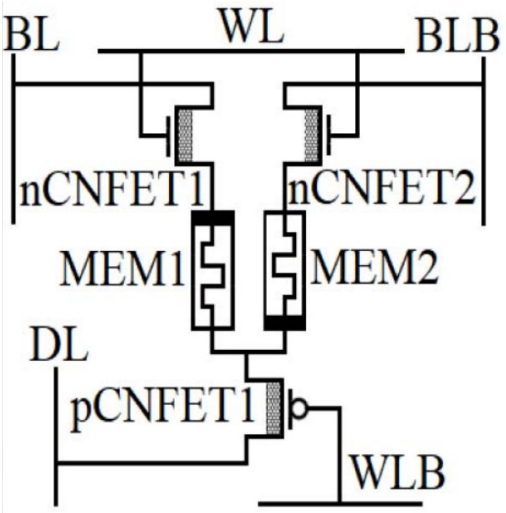


Figura 39: Cella di memoria RRAM in configurazione 3T2R

Durante l'operazione di scrittura, WL viene portata ad un valore alto per attivare i transistor di accesso nCNFET1 e nCNFET2. Al pCNFET1 si accende connettendo WLB a ground. A seconda dei dati da scrivere, BL/BLB e DL vengono caricati alla tensione di alimentazione, VDD, o scaricati a ground. Considera il caso di scrittura dello "0", ovvero di trasformare MEM1 in uno stato di bassa resistenza, R_{on}, e MEM2 in uno stato di alta resistenza, R_{off}. In questo caso, DL viene mantenuto a ground mentre sia BL che BLB vengono portati a V_{DD}. Poiché il terminale negativo (lato non drogato) di MEM1 è collegato a DL e quello positivo (lato drogato) a BL, una tensione positiva attraverso il memristor lo porta dallo stato di alta resistenza ad uno di bassa. Al contrario, MEM2 viene portato allo stato di alta resistenza per via del potenziale negativo. Per scrivere "1", DL viene mantenuto su V_{DD} mentre BL/BLB viene portato a ground. All'inizio dell'operazione di lettura, tutti gli interruttori CNFET sono attivati. Sia BL che BLB sono precaricati mentre DL è fisso a ground. Come discusso in precedenza, si considera che un memristor nello stato R_{on}/R_{off} memorizzi "0"/"1". Poiché la velocità con cui la tensione cade attraverso i memristori dipende dalle loro resistenze, entrambe le linee di bit si scaricano e si ottiene una differenza di tensione di 50 mV tra BL e BLB. Se venisse applicata una differenza di tensione costante tra i terminali di un memristor per un periodo di tempo finito, ciò potrebbe portare alla commutazione resistiva e, di conseguenza, alla perdita dei dati memorizzati. Pertanto, viene applicato un impulso di lettura di durata appena sufficiente a garantire una differenza di tensione di 50 mV tra le linee di bit. Per prevenire che ogni volta che BL o BLB si scaricano durante l'operazione di lettura, si ottiene un'effettiva caduta di tensione ai capi dei

memristor (la quale potrebbe disturbare lo stato delle celle semi-selezionate, portando alla perdita dei dati archiviati), la cella utilizza i tre interruttori CNFET per mitigare il problema dei disturbi.
Lo stato della resistenza può variare con più cicli di programmazione/cancellazione poiché una differenza di tensione tra i terminali di un memristor per un periodo di tempo sostanziale porta alla commutazione resistiva a causa dello spostamento delle cariche ioniche al suo interno. Tuttavia, ciò può essere evitato garantendo che il processo di lettura sia molto veloce. In particolare, il ritardo di lettura dovrebbe essere considerevolmente più breve della velocità di commutazione del memristor. Vale a dire, anche se ai terminali dei memristor viene applicata una tensione fissa, la commutazione non avverrà se la durata del processo di lettura è notevolmente più breve. Il tempo di commutazione di 3T2R è 47.78 volte più lungo del tempo necessario per la lettura (ritardo di lettura) con V_{DD} = 2 V. Pertanto, il processo di lettura è sufficientemente veloce da prevenire la deriva dello stato. Inoltre, la deriva dello stato dipende anche dalla durata e dal tipo di impulso di lettura utilizzato. Per garantire che i dati memorizzati rimangano intatti, le WL e WLB vengono applicati per una quantità di tempo appropriata per ottenere una larghezza di impulso sufficiente per completare con successo l'operazione di lettura senza danneggiare i dati memorizzati.
Tutti gli interruttori CNFET vengono mantenuti nello stato OFF durante la modalità standby. Sebbene non venga eseguita alcuna operazione durante la modalità di attesa, una piccola quantità di potenza di dispersione viene dissipata a causa della dispersione del gate su pCNFET1 poiché WLB viene mantenuto alla tensione di alimentazione V_{DD}.

VARIAZIONI DEL FUNZIONAMENTO

Temperatura
La temperatura gioca un ruolo chiave nel determinare le prestazioni delle celle di memoria basate su memristor. Questo perché vari parametri che definiscono il funzionamento di un dispositivo memristivo sono influenzati dalla variazione di temperatura. Ad esempio, il meccanismo fondamentale dietro l'attività di commutazione di un memristor a base di ossido è determinato dalla migrazione di ioni attivati termicamente. Inoltre, anche la velocità di diffusione, la densità dei portatori di carica e gli stati di resistenza sono influenzati dai cambiamenti di temperatura.
Lo studio dell'impatto delle variazioni di temperatura sulle caratteristiche della cella è eseguito attraverso simulazioni di scrittura e lettura variando la temperatura, in passi di cinque, da 27 ◦C a 125 ◦C a VDD = 2 V.
Nei MOSFET, l'aumento della temperatura porta a due effetti decisivi: diminuzione della tensione di soglia e aumento dello scattering. Mentre l'aumento dello scattering riduce la mobilità dei MOSFET e porta ad una diminuzione della corrente di pilotaggio, la diminuzione della tensione di soglia porta ad un aumento della corrente di pilotaggio. Il meccanismo di scattering, invece, non è dominante nei CNFET in quanto, a causa del lungo percorso libero

medio dei portatori, la loro mobilità rimane per lo più inalterata dall'incremento della temperatura. La mobilità degli ioni di ossigeno, invece, aumenta con l'aumentare della temperatura nei memristori basati su HfOx, mentre la tensione di set/reset e il rapporto di resistenza dei memristori diminuiscono. Di conseguenza, anche la velocità di commutazione aumenta con l'aumento della temperatura.

RTN

Il Random Telegraph Noise (RTN) [14] porta ad una fluttuazioni di corrente nei dispositivi RRAM, specialmente nello stato di alta resistenza (HRS). Inoltre, induce instabilità di lettura nelle celle di memoria basate su memristor. Ciò può essere attribuito al fatto che, dopo ogni ciclo di ripristino/impostazione, il filamento conduttivo (CF) viene rotto e ricostruito, il che porta ogni volta ad una struttura leggermente diversa. È pertanto necessario considerare l'impatto di RTN sulle operazioni di lettura dei circuiti basati su memristor.

L'RTN è noto per essere un fenomeno a bassa frequenza nei dispositivi CNFET ed è prevalente al di sotto di 1 KHz. Anche le fluttuazioni di corrente dovute alla RTN nei dispositivi RRAM aumentano con la diminuzione della frequenza operativa

La fluttuazione di corrente a 10 Hz è considerevolmente più alto che a 10 KHz, mentre è circa annullato ad una frequenza elevata come 1 MHz. Dato che la memoria cache è in grado di funzionare in modo affidabile in un campo di frequenza molto elevato (>> 1 MHz), RTN può essere soppresso se il dispositivo RRAM viene utilizzato nella stessa gamma. Pertanto, la cella RRAM 3T2R opera a una frequenza molto superiore a 1 MHz ed è in grado di mitigare le instabilità indotte da RTN.

MEMORIA DEL FUTURO: ULTRARAM

La UltraRAM è una delle memorie non volatili più performanti ed innovative. Esse possiedono un'alta velocità di archiviazione con una bassa energia, ed in più hanno la capacità di mantenere un dato memorizzato per lunghissimi periodi.

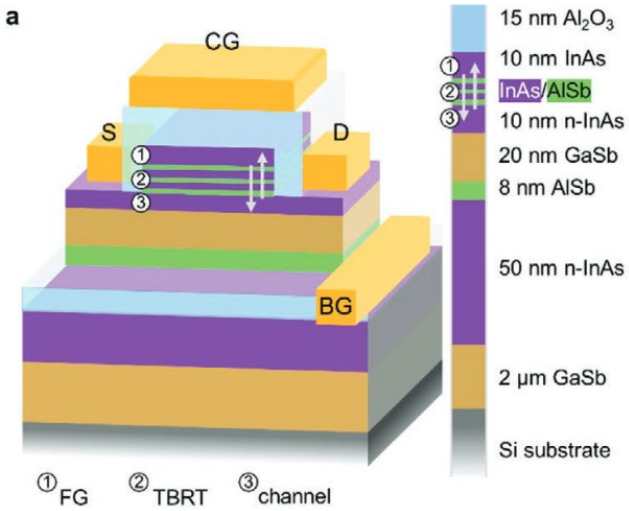


Figura 40: Composizione cella UltraRAM

UltraRAM è una memoria basata sulla carica in cui lo stato logico è determinato dalla presenza o dall'assenza di elettroni nel Floating Gate (FG). Come si può vedere nella Figura 40, l'FG è isolato elettricamente dal gate di controllo (CG) dal dielettrico Al_2O_3 [11]. La presenza di elettroni nell'FG, che definisce uno stato logico '0', esaurisce i portatori nel canale InAs di tipo n sottostante, riducendone la conduttanza. Pertanto, lo stato logico della memoria viene letto in modo non distruttivo misurando la corrente attraverso il canale quando viene applicata una tensione tra i contatti di source e drain. Il componente finale della memoria è il back-gate (BG) InAs, che consente di applicare tensioni verticalmente attraverso la pila di gate per varie operazioni. La novità alla base della memoria è la struttura TBRT (triple-barrier resonant tunneling) che, a differenza dei single-barrier, possono essere commutate da uno stato altamente resistivo elettricamente a uno stato altamente conduttivo mediante l'applicazione di soli ±2,5 V. Ciò si ottiene mediante un'attenta progettazione degli spessori delle barriere AlSb e degli strati InAs Quantum Well (QW). Quando la memoria è nello stato di ritenzione, cioè quando al dispositivo non viene applicata alcuna tensione, gli stati fondamentali degli elettroni nei QW TBRT sono disallineati tra loro. La non volatilità è rafforzata dagli stati fondamentali QW che risiedono ad un'energia insolitamente elevata per una struttura a tunnel risonante. Ciò è dovuto a una combinazione dei QW ultrasottili e della bassa mole effettiva di elettroni nell'InAs. In questo stato, il TBRT fornisce una barriera che impedisce il trasferimento di elettroni dentro o fuori dal FG. Tuttavia, l'applicazione di una polarizzazione adeguata attraverso il dispositivo, inclina la banda di conduzione in modo tale che gli stati fondamentali del TBRT QW si allineino con gli stati degli elettroni occupati nel canale, durante l'operazione di programmazione, o nell'FG, durante l'operazione di cancellazione. Ciò consente agli elettroni di muoversi

rapidamente attraverso la regione TBRT nella direzione prevista mediante il processo quantomeccanico intrinsecamente veloce del tunneling risonante. A causa delle basse tensioni richieste e della bassa capacità per unità di area del dispositivo rispetto alla DRAM, sono previsti livelli bassi di energie di commutazione dello stato logico (10^{-17} J) per memorie UltraRAM con dimensioni di 20 nm (due e tre ordini di grandezza inferiori rispettivamente a DRAM e flash).

PRESTAZIONI

Il test di ritenzione dello stato della memoria è stato effettuato a temperatura ambiente su un dispositivo con lunghezza di gate di 20 μm misurando ripetutamente I_{S-D} con una polarizzazione V_{S-D}=0.2V (in assenza di polarizzazione V_{CG-BG}). La conservazione della memoria è stata confermata per più di 24 ore sia per gli stati di programmazione che per quelli di cancellazione utilizzando più di10^6 operazioni di lettura, limitate solo dalla durata dell'esperimento.

C'è un decadimento iniziale nel contrasto I_{S-D} tra i due stati logici prima di stabilizzarsi a circa 22 μA dopo 14 ore. Per studiare ulteriormente la ritenzione della memoria, ΔI_{S-D} è stato tracciato su scala logaritmica e sono stati apportati diversi adattamenti ai dati (Figura 41).

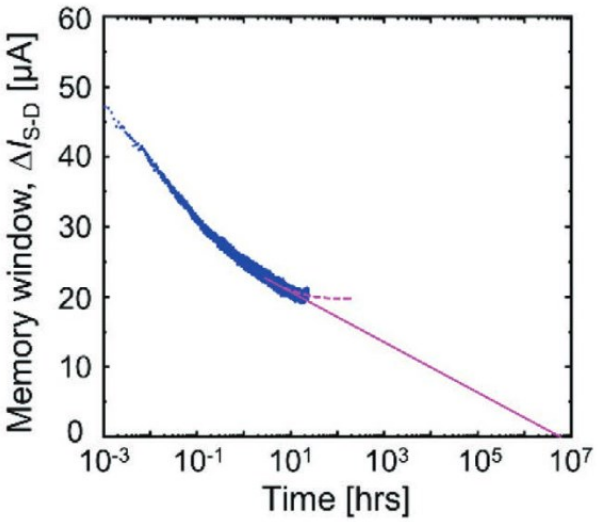


Figura 41: variazione della corrente I_{S-D}

Estrapolando queste linee adattate al punto in cui ΔI_{S-D}=0, cioè quando la finestra di memoria si chiude, è possibile stimare il tempo di ritenzione della memoria. Dopo 14 ore, è difficile la determinazione del tempo di ritenzione, come mostrato dalla linea tratteggiata nella Figura 41, che si estende all'infinito. Pertanto, viene mostrato un secondo adattamento (linea continua), che segue il decadimento degli stati di memoria, prima delle ultime 10 ore. Ciò fornisce un limite inferiore estremamente conservativo per la ritenzione della memoria di almeno 10^7 ore, ovvero più di 1000 anni.

Il test di resistenza è stato effettuato a temperatura ambiente mediante cicli di programmazione-lettura-cancellazione-lettura su un secondo dispositivo 20 μm che utilizza impulsi V_{CG-D} di durata 5 ms, rispettivamente di +2.1V e −2.55V, con misurazioni di lettura I_{S-D} a V_{S-D} = 0.2V in assenza di polarizzazione del gate. La cella di memoria è stata sottoposta con successo a 10^6 cicli di lettura-cancellazione-lettura del programma con una finestra di memoria stabile e senza degrado. La cella ha avuto zero guasti durante i cicli e <50 commutazioni parziali durante i 10^6 cicli. Di particolare rilievo è la natura riproducibile dei valori I_{SD} per gli stati programmazione e cancellazione. La deriva nell'I_{SD} è stata attribuita a un'asimmetria nel processo di programmazione/cancellazione. Successivamente è stato dimostrato dalle simulazioni che

questa deriva è il risultato dell'asimmetria della struttura TBRT, consentendo il tunneling risonante a una tensione inferiore per il ciclo di programmazione rispetto al ciclo di cancellazione. Di conseguenza, l'uso di tensioni simmetriche, come ±2.5 V, provoca una sovra-programmazione, in modo tale che in un ciclo di programma vengono aggiunti più elettroni di quanti ne vengono rimossi da una cancellazione, producendo tale deriva. Fortunatamente, lo stato si stabilizzerà rapidamente con cicli ripetuti una volta identificate le tensioni corrette.
Il test di resistenza su questo dispositivo è stato esteso di un ulteriore ordine di grandezza utilizzando una metodologia leggermente modificata per aumentare la velocità di ciclo riducendo sostanzialmente il numero di operazioni di lettura, in quanto questi richiedono molto più tempo della programmazione e della cancellazione a causa delle limitazioni delle apparecchiature di test. Il processo consiste di una serie di cicli programmazione-lettura-cancellazione-lettura e programmazione-cancellazione senza lettura, per un risultato totale di poco più 10^7 cicli di programmazione/cancellazione applicati al dispositivo.

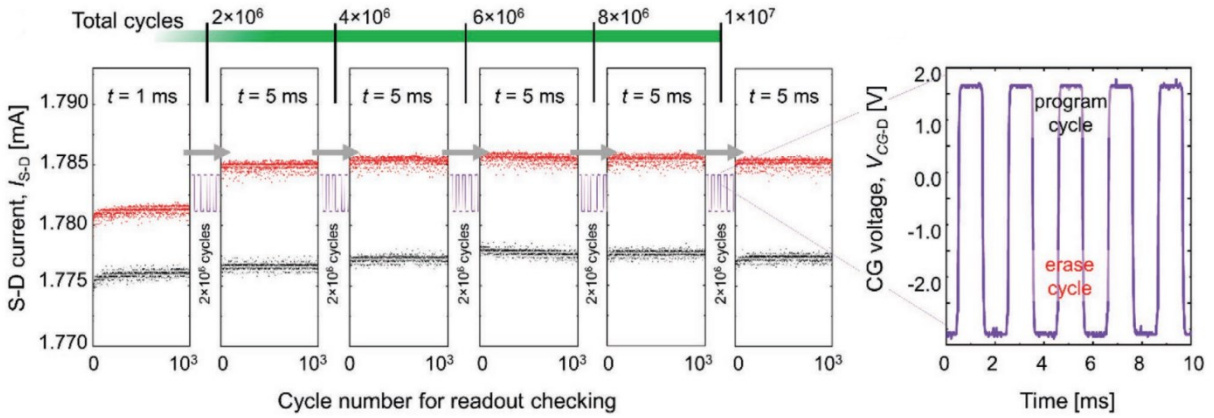


Figura 42: cicli di programmazione/cancellazione della cella UltraRAM

Come si può vedere chiaramente nella Figura 42, non vi è alcuna degradazione della finestra di memoria $\Delta I_{S\text{-}D}$ durante questi test, il che significa che la durata è almeno 10^7.

CONSIDERAZIONI FINALI

I risultati mostrano una qualità eccellente con interfacce materiali brusche e una bassa densità di difetti superficiali di $(2.1 \pm 0.1) \times 10^8$ cm^{-2}. I test sui dispositivi di memoria a cella singola fabbricati mostrano un forte potenziale, con i dispositivi che dimostrano una finestra di memoria chiara durante le operazioni di programmazione/cancellazione minori di 10 ms, che è notevolmente veloce per i dispositivi con lunghezza di gate tra 10 e 20 µm. La tensione di programmazione/cancellazione di circa 2.5V del dispositivo si traducono in un'energia di commutazione per unità di area rispettivamente 100 e 1000 volte inferiore rispetto a quelle delle DRAM e delle Flash. Tempi di ritenzione estrapolati superiori a 1000 anni e test di resistenza senza deterioramento di oltre 10^7 cicli di programmazione-cancellazione dimostrano che queste memorie sono non volatili e hanno un'elevata resistenza.

Sono in corso ulteriori lavori per migliorare la qualità epitassiale, mettere a punto il processo di fabbricazione, implementare un progetto di canale normalmente spento e ridimensionare i dispositivi.

CONCLUSIONE

In questo documento sono state analizzate diverse configurazioni delle memorie utilizzando la tecnologia CMOS per aumentare particolari prestazioni. In generale, ogni categoria di memoria ha le sue peculiarità, le quali le rendono più adatte per determinate applicazioni e meno per altre. Perciò, facendo variare la diposizione ed il numero di componenti, è possibile incrementare determinate prestazioni, mentre penalizzare altre. Per questo motivo va fatta un'attenta analisi del circuito e dell'utilizzo. È stato analizzato, come nel caso della memoria SRAM, che la variazione delle dimensioni dei transistor non possono andare sotto una determinata soglia affinché si non si abbia comportamenti imprevedibili. C'è bisogno quindi di un'innovazione delle memorie attraverso componenti innovativi.

Nella trattazione sono state analizzate diverse tecnologie che si sono poste come scopo di migliorare determinati aspetti, come la MRAM, le quali avranno sicuramente un futuro promettente e saranno sempre più richieste per cooperare e in alcuni casi sostituire le attuali SRAM e DRAM, che si ricorda essere memorie volatili. Le previsioni indicano che nel 2029 si raggiungerà un aumento dei ricavi dalla vendita delle MRAM di 170 volte rispetto al 2018, con richiesta globale da 0.1 petabyte nel 2019 a circa 1 milione di Petabyte nel 2029 [20].

Le crescenti richieste computazionali necessarie per molti miglioramenti nell'intelligenza artificiale hanno portato molti a ipotizzare che le implementazioni ReRAM potrebbero essere estremamente utile per eseguire applicazioni di AI e Machine Learning. I ricercatori della School of Engineering dell'Università di Stanford Havel hanno creato una RRAM che "esegue l'elaborazione dell'intelligenza artificiale all'interno della memoria stessa, eliminando così la separazione tra le unità di calcolo e di memoria". È due volte più efficiente dal punto di vista energetico rispetto allo stato dell'arte. La UltraRAM, invece, è ancora in fase di ricerca. Dai risultati sorti dalle simulazioni, però, si può sperare in un esito positivo. Essendo non volatile, a basso consumo energetico (mille volte inferiore ad una FLASH) e dalla robustezza eccezionale, sarà la prossima metà della tecnologia sulle memorie [11].

BIBLIOGRAFIA

[1] Bhaskar, A. (2017). *Design and Analysis of Low Power SRAM Cells.* Internationale Conference on Innovations in Power and Avanced Computing Technologies.

[2] Chua, L. (2011). *Resistance switching memories are memristors.* Springer.

[3] Hyangwoo Kim, H. C.-T.-K. (2022). *Highly Reliable Memory Operation of High-Density Three-Terminal Thyristor Random Access Memory.* SpringerOpen.

[4] Jack Dongarra, V. V. (2023). *Supercomputing Frontiers and Innovations.*

[5] Jagan Singh Meena, S. M.-Y. (2014). *Overview of emerging nonvolatile memory technologies.* Springer.

[6] Jan M. Rabaey, A. C. (2020). *Circuiti integrati digitali.* Pearson.

[7] Jaydeep P. Kulkarni, A. G. (2011). *A Read-Disturb-Free, Differential Sensing 1R/1W Port, 8T Bitcell Array.* IEEE.

[8] Nawang Chhunid, G. K. (2016). *Analysis of Performance of 3T1D Dynamic Random-Access Memory Cell.* International Journal of Electronics and Communication Engineering.

[9] Nishant Patil, A. L.-S. (2009). *Wafer-Scale Growth and Transfer of Aligned Single-Walled Carbon Nanotubes.* IEEE.

[10] Peter Bukelani Musiiwa, S. A. (2015). *Comparative Evaluation of different 3T DRAM Cells at 45nm Technology.* International Conference on Communication Networks.

[11] Peter D. Hodgson, D. L. (2022). *ULTRARAM: A Low-Energy, High-Endurance, Compound-Semiconductor Memory on Silicon.* Advanced Electronic Materials.

[12] R. Bez, E. C. (2003). *Introduction to Flash Memory.* IEEE.

[13] Richard C. Jaeger, T. N. (2018). *Microelettronica.* Mc Graw Hill Education.

[14] Soumitra Pal, S. B.-H. (2019). *Design of Power- and Variability-Aware Nonvolatile RRAM Cell Using Memristor as a Memory Element.* Electron devices society.

[15] Velguri Suresh Kumar, M. S. (2020). *Design and Implementation oh Three transistor SRAM Cell using 45nm CMOS Technology.* IEEE.

[16] William M. Regitz, J. A. (1970). *Three-Transistor-Cell 1024-Bit 500-ns MOS RAM.* IEEE Journal of Solid-State Circuits.

[17] Xiaoyao Liang, R. C.-Y. (2008). *Replacing 6T SRAMs with 3T1D DRAMs in the L1 data cache to combat process variability.* IEEE computer society.

[18] Yogesh N. Thakare, D. S. (2016). *Analysis of Power Dissipation in Design of Capacitorless Embedded DRAM.* International Conference on Automatic Control and Dynamic Optimization Techniques, Pag.: 750-755.

[19] R. Pashley, S.Lai, (1989). *Flash Memories: The Best of Two Words*, IEEE spectrum, Pag.: 30-33

[20] R. Mertens, (2019). *Analysts expect MRAM revenues to grow 170X by 2029 to reach $4billion,* Forbes

[21] (2022) *Stanford engineers present new chip that ramps up AI computing efficiency,* Stanford (online)