ALMA MATER STUDIORUM - UNIVERSITA' DI BOLOGNA

SECONDA FACOLTA' DI INGEGNERIA

CON SEDE A CESENA

CORSO DI LAUREA IN INGEGNERIA INFORMATICA

Classe 9

sede di Cesena

TESI DI LAUREA

In

ELETTRONICA DEI SISTEMI DIGITALI

NUOVE TECNOLOGIE PER L'IMPLEMENTAZIONE DI MEMORIE NON VOLATILI

CANDIDATO

Pierfrancesco Ranaldo

RELATORE:

Chiarissimo Prof. Aldo Romani

Anno Accademico 2011 / 2012

Sessione I

Indice

Introduzione

Il grande numero e l'estesa varietà di dispositivi digitali con i quali ad oggi ogni individuo può realizzare complesse e rapide elaborazioni di dati, presentano al loro interno necessariamente elementi circuitali dedicati alla memorizzazione dei dati delle applicazioni, e cioè delle memorie.

Relativamente a questi dispositivi dunque, gli sforzi attualmente sostenuti in fase di progettazione ed implementazione sono principalmente concentrati nella realizzazione di componenti che siano compatibili con le richieste di mercato, e cioè: (1) che siano in grado di concentrare grandi capacità dati in chip però sempre più piccoli; (2) che presentino bassi costi energetici per la memorizzazione delle informazioni digitali; (3) che richiedano intervalli temporali piuttosto piccoli per eseguire le fasi di lettura e scrittura dei bit; (4) che garantiscano ovviamente costi di fabbricazione non eccessivamente elevati.

Le Memorie a Semiconduttore sono gli elementi di memoria attualmente di maggior successo, vista la loro presenza nella quasi totalità dei dispositivi digitali, e cioè nei comuni dispositivi di riproduzione audio e video, o nelle fotocamere e videocamere digitali, o ancora nei personal computer, ecc. .

Esistono diversi criteri per classificare i sistemi di memorizzazione di cui sopra (vedi in base: alla modalità di accesso sequenziale/non sequenziale.; alla velocità di scrittura/lettura; alla possibilità di eseguire più volte operazioni di scrittura; al costo del dispositivo [1]), ma per la redazione di questo elaborato è stato utilizzato come elemento discriminante la proprietà di “non volatilità/volatilità”, e cioè la caratteristica del componente di memoria che indica rispettivamente se il medesimo

abbia o meno la capacità di conservare i bit precedentemente salvati anche in assenza della propria tensione di alimentazione.

In particolare, le recenti buone performance offerte dalle nuove generazioni di memorie non volatili, ha reso notevolmente convenienti l'impiego di questi elementi all'interno dei comuni dispositivi digitali, all'interno dei quali appunto ricoprono generalmente il ruolo di memoria secondaria.

In quest'ultimo contesto, le Flash Memory rappresentano i componenti di memoria maggiormente diffusi sul mercato: basti infatti pensare al sempre più frequente utilizzo come Memorie di Massa allo Stato Solido (SSD, Solid State Drive) che ne fanno sia complessi, che semplici sistemi digitali generalmente in luogo dei più lenti dischi magnetici [2].

La caratteristica principale delle Flash Memory è rappresentata dal fatto che i valori logici sono salvati iniettando o rimuovendo della carica elettrica all'interno dei gate isolati (Floating Gate) dei quali dispongono i transistori MOS a Floating Gate che costituiscono le celle di memoria di questa famiglia di dispositivi.

Tuttavia, anche gli elementi di memoria Flash presentano alcuni problemi (vedi: la modalità “a blocchi” di cancellazione dei bit salvati; il limitato numero di cicli di programmazione-cancellazione effettuabili; la possibile insorgenza di disturbi nelle celle adiacenti a quella letta; ed l'esistenza di un limite inferiore alle possibilità di scaling fissato a 19nm [2]), che spingono così la ricerca verso lo sviluppo di nuove e più promettenti tecnologie implementative di memorie non volatili.

In virtù di quanto detto precedentemente, si è pensato di articolare il seguente lavoro di tesi in cinque capitoli, che: forniscano una descrizione delle principali caratteristiche delle tecnologie implementative, attualmente maggiormente diffuse, di memorie a semiconduttore volatili e non

(Capitolo 1); espongano in modo più dettagliato quali sono i materiali utilizzati, i principi alla base del funzionamento, le implementazioni architetturali ed i problemi delle nuove memorie non volatili
Magnetoresistive (MRAM), Ferroelettriche (Fe-RAM) ed a Cambiamento di Fase (PCM) (vedi Capitoli 2,3,4); ed in ultimo, che realizzino un confronto tra le performance che caratterizzano i dispositivi di memorizzazione di cui sopra (Capitolo 5).

Capitolo 1 – Memorie a Semiconduttore.

La grande diffusione sul mercato di apparecchiature digitali portatili, ha spinto la ricerca verso lo sviluppo di dispositivi di memoria che offrano grandi capacità dati, concentrate però in chip estremamente performanti e poco costosi.

In particolare, l'esigenza di realizzare dispositivi di memoria sempre più piccoli, ha permesso alle Memorie a Semiconduttore di affermarsi pienamente nell'ambito dell'Elettronica di Consumo, essendo questa tecnologia implementativa l'unica a permettere che in un medesimo circuito, siano tra loro integrati i componenti adibiti alla memorizzazione vera e propria dei bit logici, con appunto gli elementi circuitali che garantiscono e gestiscono il funzionamento complessivo della stessa.

In realtà, all'interno della generale categoria delle Memorie a Semiconduttore possono essere distinte numerose sotto-tipologie di dispositivi, discriminabili tra loro sulla base di alcuni principali criteri, e cioè: (1) il livello di densità d'integrazione; (2) il tempo di accesso alle celle in lettura/scrittura; (3) la possibilità o meno di riprogrammare le celle in cui sono stati in precedenza salvati dei bit; (4) il numero di programmazioni che è possibile realizzare prima di riscontrare degradi di performance e l'insorgenza di rotture nel dispositivo (vedi endurance); (5) il numero di singole celle per le quali è possibile modificare lo stato di bit, senza appunto che venga alterato anche quello delle celle adiacenti (vedi granularità); ed infine, (6) la volatilità/non volatilità dei valori logici scritti.

1.1 – Breve cenno alle Memorie a Semiconduttore Volatili.

I dispositivi di memorizzazione RAM (acronimo di Random Access Memory) appartengono alla famiglia delle Memorie a Semiconduttore. In particolare, le memorie di cui sopra sono dispositivi volatili che garantiscono un tempo di accesso in lettura/scrittura uguale per tutte le celle presenti all'interno del chip (quindi indipendente dalla posizione della singola cella all'interno della memoria), dato che ciascuna di esse è accessibile adoperando opportunamente il proprio indirizzo di riga e colonna.

Questi dispositivi offrono la possibilità di leggere e programmare lo stato di bit delle loro celle, un numero di volte praticamente illimitato (nota: questa caratteristica non è attribuibile ai dispositivi non volatili).

Le memorie RAM in realtà possono essere ulteriormente classificate nelle sotto-classi DRAM (Dynamic RAM) e SRAM (Static RAM).

Nelle memorie SRAM, i valori logici dei bit sono salvati in matrici di celle implementate come semplici latch statici.

L'implementazione di questi componenti interni di memoria comporta ovviamente una maggiore occupazione di area per le celle, e ciò incide negativamente sul grado di densità d'integrazione (piuttosto basso per l'appunto [1]) offerto da questa categoria di memorie.

Al contempo però, le RAM Statiche mostrano consumi energetici piuttosto ridotti (non sono infatti richieste le “operazioni di refresh” tipiche delle DRAM, vedi sotto) ed una notevole velocità nell'esecuzione delle operazioni di accesso (in lettura/scrittura).

Proprio in virtù di quest'ultima caratteristica, ed in ragione inoltre del fatto che le stesse SRAM sono al contempo poco capienti e parecchio costose

(il processo di realizzazione dei latch su chip particolarmente piccoli non è assolutamente economico), rende queste memorie particolarmente adatte per l'implementazione di memorie Cache per le CPU dei calcolatori elettronici [2].

Per quanto riguarda invece le RAM Dinamiche, esse mostrano principalmente un elevato grado di densità d'integrazione, dato che ciascuna cella è realizzata utilizzando un solo condensatore (la carica elettrica presente sulle armature indica opportunamente il valore logico alto/basso del bit memorizzato) assieme ad un "transistore d'accesso" (esso è infatti utile per accedere alla generica cella durante le usuali operazioni di lettura/scrittura).

Il punto debole di questa tecnologia è rappresentato dal fatto che nel tempo la carica si disperde per effetto della presenza di correnti di perdita, sicché si rendono necessari operazioni di ripristino dei dati precedentemente salvati. Quest'ultimo aspetto (cioè la necessità di realizzare dei "refreshing" dei bit memorizzati), rende ovviamente le DRAM più "dispendiose" rispetto alle SRAM, per quanto riguarda i consumi energetici associabili alle due classi di dispositivi.

In ultimo, l'elevato grado di densità d'integrazione delle DRAM e una velocità seppur inferiore a quelle delle SRAM, ma tutto sommato accettabile, ed in aggiunta il minor costo di questi dispositivi rispetto a quelli tipici delle SRAM, favorisce fortemente l'utilizzo di questa tecnologia per l'implementazione di Memorie Primarie all'interno di personal computer (vedi Laptop, o Workstation) e o di Console per video-game [3].

1.2 – Memorie a Semiconduttore Non Volatili.

I più elementari dispositivi di memorizzazione a semiconduttore non volatili

attualmente esistenti, sono le cosiddette memorie a “sola lettura” o ROM, acronimo appunto di “Read Only Memory”.

Evidentemente la peculiarità principale di questa classe di dispositivi sta proprio nel fatto che le celle di memoria possono essere programmate una volta sola, e cioè durante la fase di costruzione dei singoli dispositivi ROM.

In realtà, queste memorie possono essere considerate a tutti gli effetti come dei veri e propri circuiti combinatori [4], dato che gli eventuali N bit di indirizzamento delle celle possono essere utilizzati appunto come gli ingressi delle reti combinatorie che generano le corrispondenti 2^N uscite (eventualmente formate da M bit) [4].

In particolare, siccome le memorie ROM presentano anche esse la tipica struttura matriciale riga-colonna, allora la realizzazione o meno di una connessione [generica linea delle 2^N uscite del decodificatore degli ingressi → generica linea di uscita della ROM] per mezzo di un diodo; o in alternativa, la presenza o meno di un transistore BJT NPN la cui base risulta essere comandata da una generica linea delle 2^N uscite del decodificatore degli ingressi, e con il collettore a V_{DD} e l'emettitore collegato ad una delle M uscite della ROM, determina la presenza di un valore logico rispettivamente alto o basso all'interno della corrispondente cella di memoria (nota: l'utilizzo della seconda architettura, e cioè quella a BJT, permette di realizzare dispositivi estremamente veloci in fase di lettura [4]).

Una prima evoluzione delle elementari ROM è rappresentata dalle memorie PROM (Programmable ROM), aventi anch'esse architettura matriciale, ma che presentano in corrispondenza di ogni intersezione [generica linea delle 2^N uscite del decodificatore degli ingressi → generica linea di uscita della memoria] un fusibile od un antifusibile, capaci rispettivamente di attribuire permanentemente un valore logico alto/basso

(data la loro rispettiva bassa/alta resistenza elettrica [5]) alle celle di memoria, nel caso ovviamente che i medesimi dispositivi siano ancora “non programmati” (nota: prima della fase di programmazione, quindi, tutti i bit salvati in una PROM sono a valore “1” o “0” rispettivamente).

In realtà, la “programmazione” delle celle con un valore logico pari rispettivamente a “0” od “1”, avviene applicando una sovracorrente [6] od una sovratensione [7] al fusibile ed all'antifusibile delle celle: queste operazioni sono irreversibili e pertanto i dispositivi di memoria di questo tipo sono in realtà programmabili una sola volta dall'utente.

La differenza quindi esistente tra le ROM e le PROM è che quest'ultimi dispositivi possono essere programmati anche in un istante successivo a quello in cui il medesimo è effettivamente costruito [5].

Il fatto che i valori logici delle memorie ROM e PROM possano essere scritti una sola volta, rappresenta una delle ragioni che ha limitato fortemente l'utilizzo di questi dispositivi. Nel tempo, infatti, è emersa sempre di più l'esigenza di avere a propria disposizione elementi di memoria che potessero essere programmati più volte.

In questo contesto, assume una notevole importanza lo sviluppo delle memorie EPROM (acronimo di Erasable Programmable Read Only Memory), essendo appunto questa categoria di memorie, la prima a garantire la possibilità di poter riprogrammare i bit delle celle.

In realtà, l'implementazione delle memorie EPROM è stata possibile grazie allo sviluppo e all'utilizzo dei MOSFET a Floating-Gate (vedi Fig. 1), che per l'appunto possono essere programmati in maniera reversibile.

Come mostrato nella figura della pagine seguente, ciascun MOSFET a Floating-Gate ha la caratteristica principale di possedere ben due gate, di cui: quello indicato come “Control-Gate” è accessibile dall'esterno, mentre quello che prende il nome di Floating-Gate è isolato completamente da

uno strato di Ossido di Silicio (SiO_2) che lo rende inaccessibile dall'esterno, nonché adatto per le operazioni di memorizzazione dei valori logici.

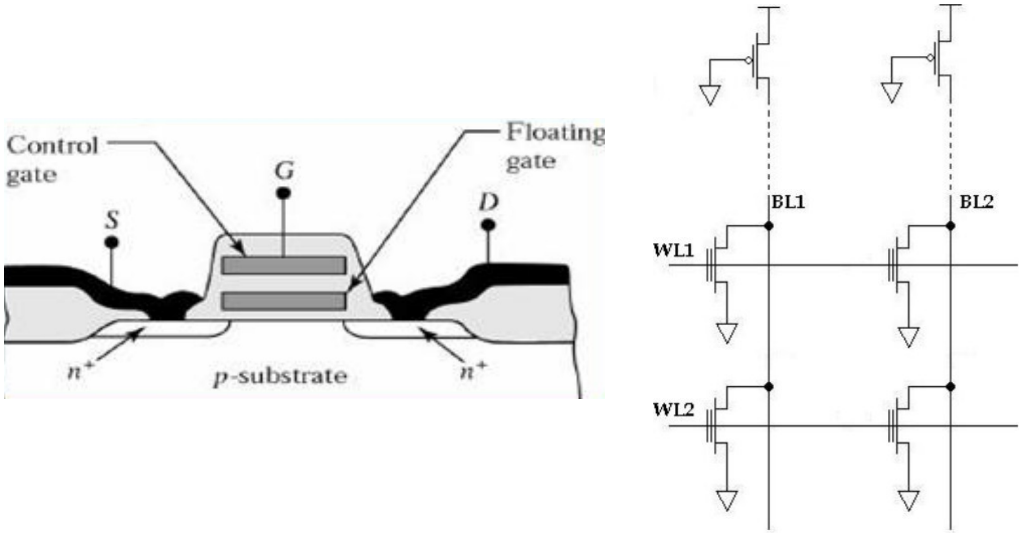


Figura1. Sezione verticale di un MOSFET a Floating-Gate utilizzato per le celle di EPROM (immagine a sinistra [8]) ed architettura a matrice di una memoria EPROM (immagine a destra [8]).

In particolare, la fase di programmazione avviene applicando elevate tensioni ai terminali di Drain e di Source che permettono di generare in corrispondenza del canale del MOSFET delle intensità di corrente così elevate, da fornire l'energia sufficiente ad una parte di elettroni per attraversare la barriera di ossido di silicio, e quindi rimanere intrappolati nel Floating-Gate per un lungo periodo di tempo.

La quantità di carica “intrappolata” sul Floating-Gate permette di modulare opportunamente la Tensione di Soglia del relativo MOSFET (si innalza il suo valore per l'appunto), e dunque lo stesso transistore risulterà acceso o spento in corrispondenza di valori di tensione applicati sul Control-Gate che saranno ben differenti da quelli utilizzati usualmente per i comuni MOSFET (nota: si tenga presente che un MOSFET-FG non programmato si comporta esattamente come un normale MOSFET).

Per quanto riguarda invece l'operazione di cancellazione, essa è totale e consiste nell'esposizione del die dell'EPROM [10] a radiazioni ultraviolette (quest'ultime infatti, forniscono l'energia necessaria agli elettroni intrappolati nei F-G per poter riattraversare la barriera isolante e abbandonare il gate isolato).

La tecnologia EPROM è attualmente ormai obsoleta, nonché è stata completamente rimpiazzata dai dispositivi EEPROM e dalle diffusissime Memorie Flash, entrambe appunto cancellabili elettricamente.

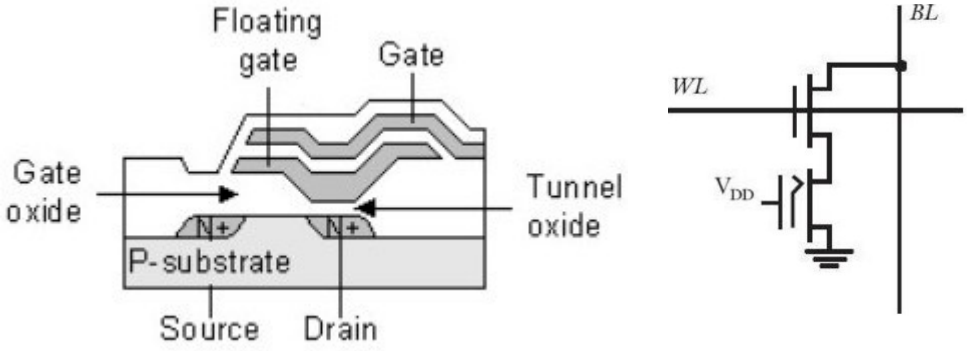


Figura 2. Sezione verticale di un MOSFET a Floating-Gate utilizzato per le celle di EEPROM (immagine a sinistra [9]) ed architettura a matrice di una memoria EEPROM (immagine a destra [8])

I dispositivi di memoria EEPROM (Electrically EPROM) si differenziano dalle EPROM per tre principali ragioni, e cioè: (1) utilizzano un particolare MOSFET a F.G., detto appunto FLOTOX [8,9] (vedi Fig. 2, a sinistra); (2) ciascuna cella presenta anche un "transistore di accesso"; (3) utilizzano un principio fisico di meccanica quantistica noto come "Effetto Tunnel di Fowler-Nordheim" [8,9] per la scrittura/cancellazione delle celle di memoria (si tenga presente che quello adottato dalle EPROM invece, è indicato come "Effetto di breakdown a valanga" o "Iniezione di elettroni caldi" [10]).

La caratteristica primaria di ogni FLOTOX è quella di presentare in corrispondenza di una regione limitata dell'area di Drain del MOSFET a F-G, una riduzione dello spessore dell'Ossido di Silicio che isola appunto il

F-G dal resto del dispositivo.

Dunque, grazie a questa peculiarità, l'iniezione di elettroni all'interno del F-G attraverso la regione [Drain → Ridotto Spessore di SiO_2], avviene pilotando opportunamente i valori di tensioni sui terminali di Control-Gate, Source e Drain (i primi due saranno a valori alti di tensione, mentre l'ultimo sarà a massa).

Contrariamente, impostando un elevato valore di tensione sul Drain e ponendo il C-G ed il Source a massa, si realizza il processo di tunnelling in maniera inversa, sicché il Floating-Gate viene svuotato dei suoi elettroni.

In realtà il processo di cancellazione è piuttosto delicato, dato che un eccessivo “svuotamento” del FG potrebbe implicare la “creazione di lacune” all'interno del gate isolato. A tal riguardo, e cioè per fare in modo che tale criticità non si verifichi mai, si dispone in corrispondenza di ogni cella di EEPROM un MOSFET di accesso (esso ovviamente è utilizzato anche in fase di lettura).

La presenza di un MOSFET di accesso implica una maggiore occupazione di area per singola cella di EEPROM rispetto appunto a quelle tipiche delle Memorie Flash.

A ciò inoltre bisogna aggiungere un valore di Endurance e di Ritenzione dei Dati per le EEPROM che è tipicamente inferiore sempre rispetto a quelli delle più moderne Flash Memory [11].

1.3 – Flash Memory.

I dispositivi di Flash Memory costituiscono l'evoluzione delle memorie EEPROM. In particolare, anche questa classe di memorie è cancellabile e riprogrammabile elettricamente esattamente come le più vecchie memorie EEPROM ma, al contrario di quest'ultime, esse presentano numerosi altri

aspetti che le hanno rese estremamente diffuse.

A tal proposito, due delle peculiarità principali delle Memorie Flash sono: (1) l'alto grado di densità d'integrazione (vedi implementazione a NAND [12]); (2) la possibilità di poter realizzare la fase [cancellazione → riprogrammazione] su blocchi o pagine di memoria (in realtà anche su singole word, vedi implementazione a NOR) [12], contrariamente a quanto invece accade per le EEPROM, per le quali appunto l'operazione di cancellazione è realizzabile su singoli byte [12], nonché risultano essere molto più lente rispetto alle Flash Memory.

Anche alla base di questa tecnologia implementativa di memoria vi sono i MOSFET a Floating-Gate, ma di una particolare tipologia indicata con il nome di ETOX [8,9].

I transistori ETOX sono utilizzati perché uniscono i vantaggi dei processi di scrittura delle EPROM a quelli di cancellazione delle EEPROM, e cioè: se per intrappolare nel F-G di un ETOX si utilizza il principio fisico del “Hot Carriers Injection”, tipico appunto della prima classe sopra citata, per svuotare il gate flottante si usa al contrario l’”Effetto di Fowler-Nordheim” delle EEPROM (vedi Fig. 3).

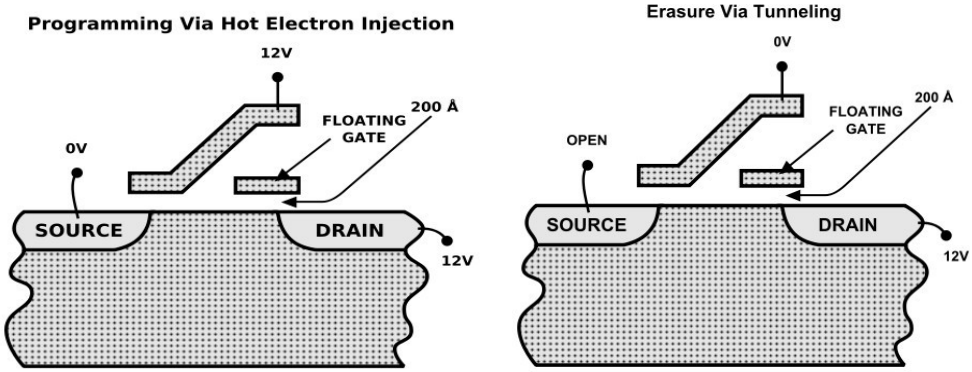


Figura 3. Sezione verticale di un transistore ETOX durante la fase di programmazione a “Hot Carriers Injection” (immagine a sinistra [12]) e sezione verticale di un transistore ETOX durante la fase di cancellazione con “Effetto Tunnel di Fowler-Nordheim” (immagine a destra [12]). Si noti come in entrambe le figure sia mostrato lo spessore “ridotto” dell'Ossido di Silicio (pari a 200Å cioè 200X10 $^{-10}$m = 20nm) in corrispondenza del terminale di Drain.

In virtù di quanto detto di sopra, l'operazione di programmazione avviene applicando una tensione piuttosto elevata sia al Control-Gate, così appunto da attivare il canale n del MOSFET (nota: si suppone di utilizzare MOSFET a FG a “canale n”), sia al terminale di Drain (vedi immagine sinistra Fig. 3).

A questo punto, mantenendo a massa il terminale di Source, allora l'elevata corrente di canale permetterà agli elettroni con sufficiente energia di attraversare la barriera di SiO_2 e di rimanere così intrappolati all'interno del Floating-Gate.

Il processo di cancellazione avviene invece, come appunto mostrato nell'immagine a destra della Fig. 3, mantenendo aperto il terminale di Source ed a massa quello di CG, sicché l'applicazione di una tensione particolarmente elevata al terminale di Drain “attira” fuori gli elettroni precedentemente “intrappolati” attraverso appunto la regione di barriera di Ossido di Silicio a minor spessore (vedi cancellazione EEPROM).

Le operazioni di programmazione e cancellazione di cui sopra sono eseguite in modalità differenti nelle Memorie Flash, dato che i dispositivi attualmente disponibili sono implementati alternativamente con “Architettura a NOR” o con “Architettura a NAND”.

E cioè, come mostrato in Fig. 4 (immagine superiore), nel caso di struttura interna a NOR, ogni cella è implementata con un ETOX avente un terminale connesso alla Bitline e l'altro a massa. Dunque, l'attivazione della Wordline, connessa per l'appunto al Control-Gate del generico ETOX di cella, permetterà o meno di accendere lo stesso MOSFET e conseguentemente di scaricare o meno la tensione di BL (appunto pari a V_{DD}), a seconda che rispettivamente nel Floating-Gate non siano stati/siano stati intrappolati degli elettroni (la BL rimane ad “1” se quindi la cella è a “0”).

Le Memorie Flash NOR sono utilizzate soprattutto per la sostituzione delle più vecchie EEPROM, dato che esse offrono ottime prestazioni nell'accesso ai dati in lettura (le Flash NOR sono impiegate generalmente per memorizzare dati che verranno raramente modificati, e che altresì subiranno letture frequenti, vedi: sistemi operativi o firmware di periferiche digitali [13]).

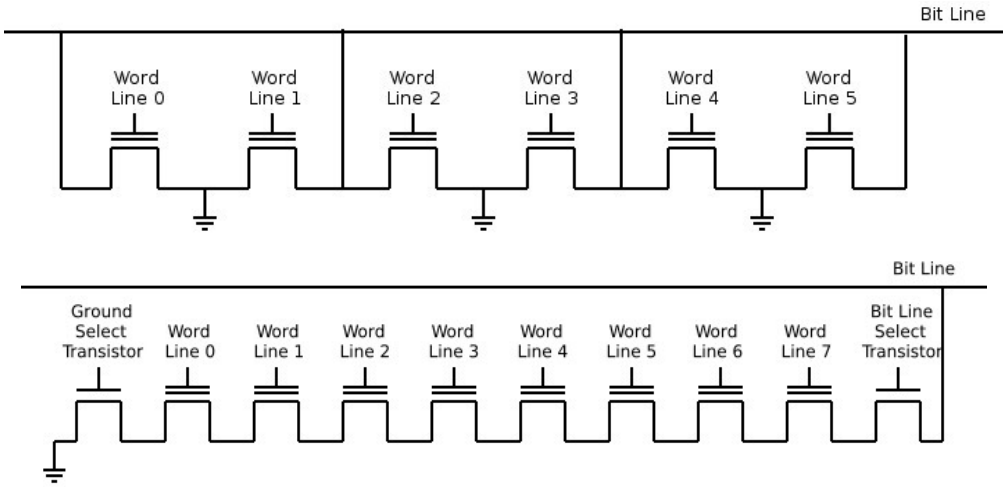


Figura 4. Architettura a NOR (immagine superiore [12,13]) ed a NAND (immagine inferiore [12,13]) di una Memoria Flash.

Per quanto riguarda invece le Memorie Flash a NAND (vedi Fig. 4, immagine inferiore), l'architettura interna è implementata connettendo in serie tra loro gli ETOX, e facendo in modo che un terminale della serie sia connesso alla generica BL, mentre l'altro a massa.

La lettura avviene portando tutte le WL, meno quella della cella da leggere, ad un valore di tensione tale da consentire l'accensione dei propri corrispondenti ETOX anche qualora i gate flottanti fossero stati programmati. Contemporaneamente, la WL della cella da leggere è portata ad un valore di tensione pari a quello di soglia di un ETOX “non programmato”, e dunque: se il medesimo MOSFET a FG si accende (nel suo FL cioè non è presente alcuna carica intrappolata), allora la serie consente alla BL di scaricarsi; se invece lo stesso ETOX non si accende,

allora la BL rimane alla tensione V_{DD}.

Il successo di quest'architettura è data dalla possibilità di realizzare, rispetto alle NOR, un processo di cancellazione più complesso e mirato, e cioè: sapendo che i “settori di cancellazione” delle FLASH-NOR e delle FLASH-NAND sono rispettivamente pari in genere a 64Kb e 8Kb [13], allora nel primo caso, la riscrittura anche di un solo byte comporta la cancellazione di 64Kb di memoria, seguita poi dalla riscrittura per intero (con le relative modifiche) del medesimo blocco; nel secondo caso, invece, le operazioni di cancellazione ed aggiornamento interesseranno solamente 8Kb di memoria.

Alla luce di ciò, si capisce che i processi di “aggiornamento dei dati” per una FLASH-NAND occupano intervalli temporale nettamente inferiori rispetto a quelli di una FLASH-NOR.

Inoltre è possibile notare come l'area complessivamente occupata dalle celle di una FLASH-NAND sia decisamente più piccola in confronto a quella delle celle di FLASH-NOR [13], dato appunto che nelle FLASH-NAND l'elevato numero di connessioni verso massa e di linee di BL [12] (tipiche appunto delle implementazioni a NOR) non sono necessari.

In virtù di ciò, la realizzazione di chip di FLASH-NAND consente di concentrate maggiore capacità dati in uno stesso chip avente inoltre un costo più contenuto rispetto a quello di un chip di FLASH-NOR [13].

I problemi principali associabili alle Memorie Flash sono principalmente tre, e cioè:

1. L'impossibilità di realizzare operazioni di riscrittura di una cella ETOX, se non dopo un ciclo di cancellazione complessiva del blocco di appartenenza della stessa cella. Ovvero, l'aggiornamento dei dati di memoria prevede l'iniezione di elettroni in tutti i MOSFET-FG del blocco,

quindi lo svuotamento dei FG per le celle che lo richiedono. A questo punto, se si volesse iniettare nuovamente degli elettroni nel FG di una o più celle che hanno subito in precedenza lo svuotamento, allora sarebbe necessario ripetere il procedimento appena descritto (cioè far ripartire il processo di “iniezione totale” e “svuotamento selettivo”).

2. La possibilità di poter eseguire sulle celle di memoria solamente un numero limitato di cicli di lettura/scrittura, prima appunto che il dispositivo presenti un degrado delle performance ed eventuali malfunzionamenti;

3. La necessità di eseguire dei cicli di ripristino dei valori logici di cella per le FLASH-NAND, dato che l'esecuzione nel tempo di numerosi cicli di lettura, porta in genere ad un'alterazione dello stato di bit delle celle adiacenti a quelle lette.

4. Impossibilità di realizzare ulteriori miniaturizzazioni dei dispositivi attualmente disponibili (vedi andare oltre i 45nm per le FLASH-NOR ed i 22nm per le FLASH-NAND [15]), poiché: (a) lo strato di SiO_2 non può raggiungere spessori inferiori ai 10nm, dato che altrimenti sarebbe fortemente a repentaglio l'affidabilità dei dispositivi di memoria (si tenga presente, infatti, che nel tempo l'esecuzione di cicli di lettura/scrittura potrebbero causare nello strato isolato, così poco spesso, l'insorgenza di percorsi di leakage per la carica del FG [14]); (b) un eccessivo scaling produrrebbe il manifestarsi di possibili fenomeni di interferenza tra celle adiacenti [15].

Per questa serie di motivi, gli sforzi dell'industria elettronica si stanno maggiormente concentrando sulla ricerca di nuove tecnologie implementative di memoria, che appunto consento di realizzare dispositivi: non volatili; aventi processi di lettura/scrittura con performance prossime a quelle delle Flash e/o delle DRAM; con livello di endurance praticamente

illimitato (proprio come le memorie volatili); ed in ultimo, aventi costi energetici e di produzione non eccessivamente elevati.

Capitolo 2 – Memorie Magnetoresistive.

Precisazioni.

Il seguente capitolo è stato scritto riassumendo e traducendo in parte l'articolo [1], dal quale inoltre sono state tratte anche le figure ed i grafici che ivi compaiono.

La tecnologia implementativa delle RAM Magnetoresistive (MRAM) combina aspetti tipici della tecnologia spintronica, con quelli della comune microelettronica dei dispositivi a semiconduttore.
I principali attributi di una MRAM sono la non volatilità dei dati in essa salvati, quindi la possibilità di eseguire un numero illimitato di cicli di scrittura sulle proprie celle, a differenza di quanto invece accade per i comuni dispositivi non volatili a floating-gate.
In particolare, l'utilizzo all'interno delle MRAM delle giunzioni a tunnel magnetico (MTJ), ha reso possibile l'implementazione di dispositivi di memoria con alto grado di densità d'integrazione, nonché particolarmente veloci, favorendo così la commercializzazione di questa famiglia di componenti.
Le prime memorie funzionanti sulla base di effetti magnetoresistivi (anni '80) [10-12], sfruttavano il fenomeno per cui la resistenza di un materiale ferromagnetico dipende dall'angolo compreso tra la direzione del campo magnetico applicato, e la corrente che nello stesso viene fatta scorrere [2].

Questo effetto prende il nome di Magnetoresistenza Anisotropica (AMR, Anisotropic Magnetoresistance) [2].

Tuttavia, memorie che utilizzano questa caratteristica della fisica dei materiali ferromagnetici, non hanno avuto grande successo commerciale, poiché presentano: (1) un basso grado di densità d'integrazione; (2) limitatissime variazioni del valore di resistenza (ciò rende per l'appunto molto complesse le operazioni di lettura/scrittura dei bit); (3) elevati consumi energetici.

In modo del tutto alternativo alle memorie a tecnologia AMR, l'utilizzo di pellicole metalliche multistrato separate da un ulteriore layer di materiale non-ferromagnetico (cioè un materiale che possiede intrinsecamente un trascurabile campo magnetico, vedi: il rame, l'alluminio o la plastica [3]), consente di sviluppare più elevati e robusti valori di magnetoresistenza, sfruttando appunto il fenomeno fisico di Magnetoresistenza Gigante (GMR, Giant Magnetoresistance) [13-16].

In particolare, la resistenza di un materiale a più strati dipende dalla direzione di magnetizzazione che gli stessi layer assumono vicendevolmente.

L'effetto GMR produce più nette variazioni del valore di resistenza della cella (nell'intervallo di 4% - 8%) rispetto a quelle mostrate per i dispositivi AMR, nonché la bassa resistività che caratterizza le pellicole multistrato produce segnali piccoli/contenuti, rendendo così compatibili questi dispositivi con quelli a semiconduttore.

In definitiva, però, una generica cella GMR presenta bassissimi valori di resistenza, e questo è fondamentalmente il fattore per cui la tecnologia in esame ha avuto poco impiego nell'ambito della realizzazione di memorie.

Lo sviluppo di materiali con giunzioni a tunnel magnetoresistivo (MTJ, Magnetoresistance Tunnel Junction) [17-19], predisposti cioè per il supporto di effetti di tunnel magnetoresistivo (TMR, Tunnelling Magnetoresistance), hanno consentito la realizzazione di MRAM con celle di memoria che mostrano maggiori valori di resistenza e cambiamenti più marcati di questo parametro.

In particolare, il materiale utilizzato per la realizzazione di un generico MTJ, è costituito da due strati esterni di materiale ferromagnetico, separati da un layer isolante (detto barriera del tunnel), tipicamente in ossido di alluminio. La corrente scorre perpendicolarmente al piano del materiale, utilizzando un fenomeno fisico di meccanica quantistica che consente agli elettroni della corrente elettrica di attraversare lo strato di barriera isolante.

La realizzazione di MRAM con tecnologia MTJ è quella sicuramente di maggior successo, dato che questi materiali permettono di realizzare celle aventi: ridottissime dimensioni (vista la loro geometria current-in-perpendicular); più elevati valori di resistenza rispetto a quelli tipici delle celle in materiale con effetto GMR, e che inoltre possono essere opportunamente dimensionati in conformità con il circuito all'interno del quale le stesse giunzioni sono adoperate [20-22].

Gli aspetti ancora da migliorare per questa tecnologia riguardano: l'uniformità del valore di resistenza delle celle (tutte devono possedere lo stesso valore di resistenza); l'integrazione degli MTJ con i circuiti CMOS; ed in ultimo, le operazioni di scrittura, le quali appunto non devono realizzare il salvataggio di valori logici spuri [23-25].

2.1 – Cella MRAM con un MTJ ed un unico transistore di isolamento.

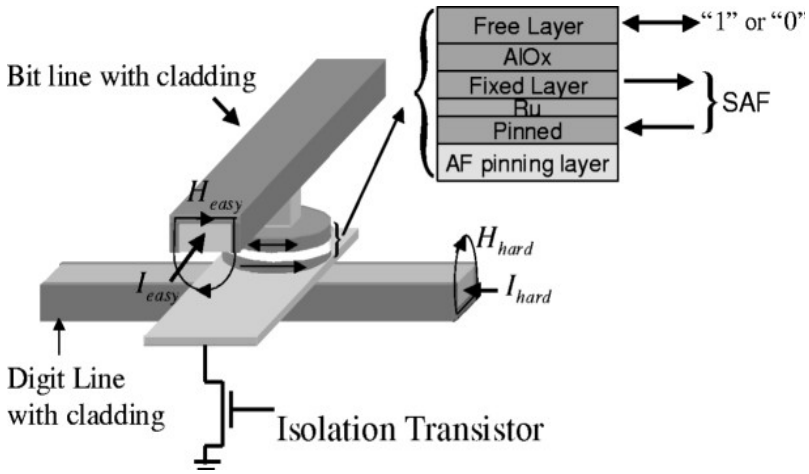


Figura 2.1. [1] Struttura di una cella di MRAM e del corrispondente MTJ in essa utilizzato (architettura 1T1MTJ). Nella figura sono anche visibili i campi magnetici di easy-axis e di hard-axis prodotti dalla Bitline e dalla Digitline, necessari appunto alla scrittura della cella.

Come mostrato in Fig. 2.1, ogni cella di memoria di MRAM a MTJ è collegata ad una coppia di linee conduttrici disposte perpendicolarmente tra loro, e di cui una collocata sopra la giunzione, e l'altra invece al di sotto della stessa.

Nella figura è possibile inoltre rilevare la presenza (in corrispondenza della base della giunzione) di un transistore che consente di isolare o meno la cella dal circuito di lettura degli stati logici della MRAM.

La corrente che scorre nelle linee che passano al di sopra ed al di sotto del MTJ quando il transistor è spento, consentono l'instaurazione dei campi magnetici necessari al cambiamento del valore di resistenza della cella, quindi ad impostare il valore logico da salvare.

In particolare, in questa implementazione la linea superiore della cella (indicata in Fig. 2.1 come Bitline) è utilizzata per l'instaurazione del "campo

magnetico di easy-axis": il verso con cui la corrente scorre all'interno di questo collegamento è fondamentale per la definizione del "valore logico" da memorizzare nel "layer libero" del MTJ della cella.

Al contrario, la connessione inferiore (indicata in Fig. 2.1 come Digitline) permette l'instaurazione del campo magnetico di hard-axis, che risulta fondamentale per ridurre la corrente necessaria alla realizzazione del processo di commutazione del valore di bit.

Come mostrato in Fig. 2.2, infatti, la assenza/presenza della corrente di Digitline comporta un corrispondente aumento/diminuzione del valore di intensità corrente di Bitline necessaria per la definizione dei valori di resistenza del MTJ.

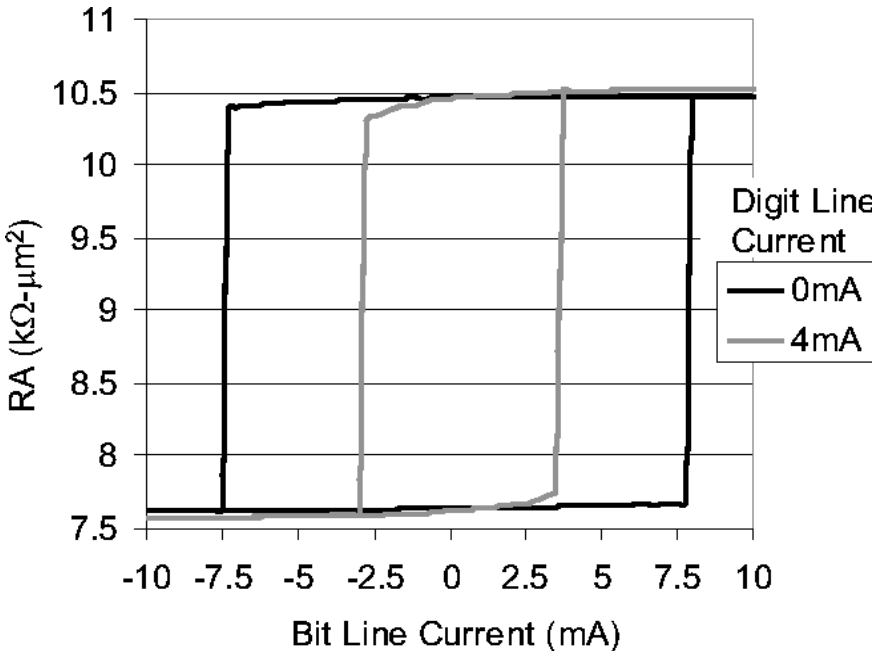


Figura 2.2. [1] Resistenza (per Area), definita in funzione dell'intensità di corrente della Bitline (caso di un MTJ 0.6 μm x 1.2 μm con 350-mV come voltaggio di riferimento). Dal confronto dei due grafici, è evidente che con la presenza della corrente di Digitline (curva non in grassetto), i valori di corrente di Bitline necessari per lo switching in resistenza, si riducono sensibilmente rispetto al caso in cui sia assente la corrente di Digit Line (curva in grassetto).

Si tenga presente che la corrente che circola all'interno della Digitline è in realtà unipolare, dato che il segno del campo di hard-axis non influenza il valore logico del bit da salvare.
Inoltre, solo in fase di lettura, il transistor risulta acceso così da poter rilevare la piccola corrente in uscita dal MTJ (circa 10 μA) e determinare così il valore di resistenza della stessa cella.
Questo tipo di implementazione è detta *"1T1MTJ"* (*OneTransistor-OneMTJ*), dato che ogni cella di memoria è costituita da un MTJ e da un singolo transistore di isolamento.
Quella 1T1MTJ è l'architettura più comune, anche se in realtà è possibile implementare dispositivi MRAM utilizzando anche altri modelli, come appunto quello a 2MTJ x Cella (offre maggiore velocità, ma minore densità d'integrazione) [26]; o quello senza transistori di isolamento [27](presenta un aumento della densità d'integrazione, ma velocità ridotte [28]).

2.2 – Materiali utilizzati per la costruzione degli MTJ.

Come già detto in precedenza, lo stack MTJ è formato da due strati di materiale ferromagnetico, separati da un sottilissimo layer in materiale isolante.
Il fenomeno di tunnelling magnetoresistivo può essere compreso se si fa riferimento al modello a due bande, ovvero: gli elettroni presenti nei due layer esterni della giunzione, possono appartenere in modo esclusivo o alla banda di spin-up, o a quella di spin-down, alle quali appunto sono in genere associate due differenti densità degli stati (D.O.S., Density of States) in corrispondenza dell'Energia di Fermi.

Si tenga presente che per “spin di un elettrone” si intende il momento angolare quantistico derivante dalle rotazioni che tale particella compie su se stessa [4]. In particolare, il fatto che un elettrone ruoti su stesso attorno ad un asse che idealmente lo attraversa, genera un corrispondente momento di dipolo magnetico (proprio in virtù del teorema di equivalenza di Ampère), che potrà essere allineato o non, con un eventuale campo magnetico esterno [4].

Inoltre, con l'espressione “densità degli stati ad un dato livello di energia” si intende il numero di stati disponibili (nonché occupabili dagli elettroni) ad una data energia presenti nel materiale allo stato solido che si considera [5].

Solitamente un normale processo di tunnelling avviene non invertendo lo spin dei singoli elettroni (in questi casi si parla infatti di spin-flip scattering), cosicché ciascuno di essi con spin up/down passa da un elettrodo ad un altro, conservando questa caratteristica.

Quando le singole magnetizzazioni dei layer più esterni sono in parallelo tra loro (vedi Fig. 2.3a), allora gli elettroni appartenenti alla banda maggiore di spin up/down attraversano il tunnel, raggiungendo rispettivamente la banda maggiore di spin up/down dell’altro elettrodo, e ciò ovviamente si verifica anche per gli elettroni che sono in banda minore (appunto di spin down/up).

Al contrario, nel caso di allineamento antiparallelo delle magnetizzazioni degli strati ferromagnetici, gli elettroni che sono in banda maggiore/minore attraversano il tunnel, andando però a finire in bande che sono rispettivamente minori/maggiori per l’elettrodo di arrivo. In questo caso, è quindi evidente che per gli elettroni appartenenti ad una stessa banda, nell’elettrodo di arrivo sono disponibili un minor numero di stati: ciò comporta una riduzione della corrente di tunnelling, che coincide con un

aumento del valore di resistenza della giunzione rispetto a quella evidenziata dalla stessa nel caso di allineamento parallelo (Fig. 2.3b).

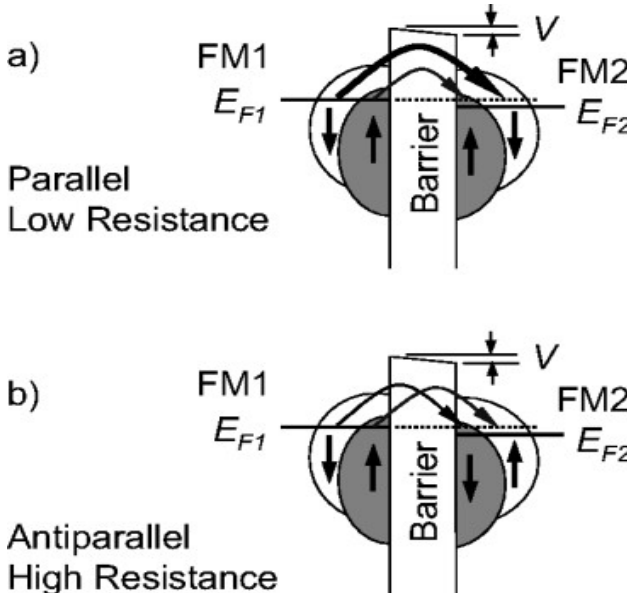


Figura 2.3. [1] Nella sezione (a), cioè nel caso di eff. di tunnelling forte, è mostrata con una freccia in grassetto la corrente degli elettroni di banda maggiore, mentre con una freccia sottile la corrente degli elettroni in banda minore. In entrambe le sezioni, sono indicate con FM1 e FM2 i due strati ferromagnetici della giunzione, per le quali: l'area in bianco rappresenta la popolazione di elettroni in banda maggiore, mentre l'area scura individua quella degli elettroni in banda minore. Infine, con V è indicato il voltaggio di riferimento (Bias Voltage) applicato alla giunzione, mentre il simbolo di E_F rappresenta l'energia di Fermi.

Detto ciò, è indicato con il nome di "Rapporto di Magnetoresistenza di Tunnel", o più sinteticamente TMR-Ratio, il valore che si ottiene mettendo per l'appunto a rapporto la differenza di resistenza evidenziata nei due stati parallelo/antiparallelo, con il valore di resistenza nel caso di allineamento parallelo (vedi relazione di sotto).

$$TMR-Ratio=\frac{R_{antiparallelo}-R_{parallelo}}{R_{parallelo}};[6]$$

Maggiori valori di TMR-Ratio (e quindi più nette differenze tra i valori di $R_{antiparallelo}$ e di $R_{parallelo}$) emergono ovviamente nei materiali che presentano

più evidenti squilibri tra la D.O.S. all'E.F. per la banda maggiore, e quella per la banda minore nei propri layer ferromagnetici.

E cioè, partendo dalla seguente relazione:

$$P=\frac{DOS(Ef)\uparrow - DOS(Ef)\downarrow}{DOS(Ef)\uparrow + DOS(Ef)\downarrow};[1,6]$$

che consente per l'appunto di calcolare la polarizzazione di spin P di un generico layer ferromagnetico prendendo in esame la DOS degli elettroni con spin-up (↑) (sono quelli che hanno orientamento di spin concorde con la direzione orientata del campo magnetico esterno utilizzato per la commutazione magnetica del valore logico di cella [6]), e quella degli elettroni con spin-down (↓) (hanno orientamento di spin discorde con la direzione orientata dello stesso campo magnetico [6]), ed utilizzando pertanto questa relazione per il calcolo delle polarizzazioni di spin dei layer FM1 ed FM2 (grandezze indicate rispettivamente con P1 e P2), è possibile determinare l'ampiezza della differenza relativa TMR che caratterizza la giunzione attraverso l'equazione:

$$TMR-Ratio=\frac{2*(P1P2)}{1-(P1P2)}.[6]$$

Si tenga presente che il calcolo diretto dei valori di resistenza per il caso di allineamento parallelo/antiparallelo è piuttosto complesso per essere descritto sinteticamente in questo capitolo.

Tuttavia, a tal riguardo è possibile affermare semplicemente che il valore di resistenza per il caso parallelo/antiparallelo è proporzionale rispettivamente a $[DOS(Ef)\uparrow_1 * DOS(Ef)\uparrow_2 + DOS(Ef)\downarrow_1 * DOS(Ef)\downarrow_2]^{-1}$ [7] ed a $[DOS(Ef)\uparrow_1 * DOS(Ef)\downarrow_2 + DOS(Ef)\downarrow_1 * DOS(Ef)\uparrow_2]^{-1}$ [7].

Infine, la differenza di potenziale presente in corrispondenza della giunzione (vedi Fig. 2.3), è fondamentale perché altrimenti il fenomeno di

tunnel degli elettroni si svilupperebbe in entrambe le direzioni, quindi gli effetti generati si compenserebbero vicendevolmente e la corrente di giunzione risultante sarebbe conseguentemente nulla. Con l'applicazione, invece, di un voltaggio di riferimento si spinge gli elettroni a muoversi verso il layer a potenziale maggiore.

In realtà i disegni di Fig. 2.3 semplificano molto il reale fenomeno di tunnelling, anche in virtù del fatto che la barriera isolante ricopre un ruolo non poco rilevante nella determinazione dell'ampiezza e del segno dello stesso effetto di attraversamento degli elettroni [29], e che il valore di TMR-Ratio dipende da alcune caratteristiche proprie dei materiali della giunzione e dal voltaggio di riferimento (è possibile registrare una riduzione del parametro TMR-Ratio in corrispondenza di un incremento del voltaggio della giunzione, supponendo che gli elettrodi siano realizzati in leghe di Nichel-Fero-Cobalto e che siano separate da una barriera isolante in ossido di alluminio).

Tipicamente, una cella di MRAM implementata con MTJ, ha un layer ferromagnetico libero (Free Layer), e l'altro invece fisso (Fixed Layer): la magnetizzazione di quest'ultimo non può essere alterata (cioè non può compiere rotazioni) con la presenza di un campo magnetico esterno, grazie al fatto che lo stesso "strato fisso" è accoppiato con altri ulteriori layer.

In particolare, la costruzione di un layer fisso che conservi nel tempo la propria direzione di magnetizzazione, può avvenire o utilizzando un materiale ad alta coercività (sarà necessario applicare un campo magnetico inverso di elevata intensità per annullare la magnetizzazione di saturazione dello stesso layer fisso), oppure accoppiando lo stesso layer con un ulteriore strato, questa volta però in materiale antiferromagnetico (si noti in Fig. 2.1, il layer fisso è disposto al di sopra di uno strato in

rutenio, il quale garantisce un forte accoppiamento dello stesso layer considerato, con lo strato antiferromagnetico indicato con il nome di "pinned-layer" [30]).

L'architettura "Free-Fixed Layers" degli MTJ, implica che le operazioni di scrittura consistano nella possibilità di invertire la sola direzione orientata di magnetizzazione del "layer libero" di giunzione.

In ultimo, la realizzazione di MRAM con celle a MTJ aventi valori di resistenza utilizzabili, prevede che lo spessore della barriera isolante sia dell'ordine degli 1.5 nm o meno.

A tal riguardo, è importante precisare che il valore della resistenza di tunneling è fortemente influenzata dallo spessore della barriera, ed in particolare esiste una dipendenza esponenziale tra resistenza e spessore del tunnel, per cui anche piccolissime variazioni dello spessore della barriera comportano grandi cambiamenti del valore di TMR [31].

Una delle sfide nell'attuale realizzazione di MRAM con celle MTJ è rappresentata dalla riduzione delle dimensioni di cella, cercando però di mantenere costante (o al più decrementando) il valore di Resistenza x Area (RA) della stessa: è fondamentale infatti che la singola cella non sviluppi resistenze eccessivamente elevate, che incidano poi negativamente sui ritardi RC delle linee circuitali.

2.3 – Operazioni di lettura e di scrittura per una cella di MRAM.

Il passaggio di corrente all'interno delle linee ortogonali (vedi Fig. 2.1) consente la programmazione dei "layer liberi" degli MTJ che formano le celle di memoria.

Il layer libero di un MTJ ha generalmente la forma allungata, cosicché l'anisotropia magnetica di forma (la magnetizzazione non sarà uguale in tutte le direzioni, ma si avrà uno o più assi preferenziali, quali appunto quelli indicati come easy-axisis [8]) consente la definizione di una barriera di energia Eb, funzionale nell'ostacolare la magnetizzazione dello stesso strato nella direzione orientata opposta a quella posseduta correntemente dal layer libero.

La programmazione della cella avviene quindi adoperando un campo magnetico avente direzione orientata coincidente con l'easy-axis (vedi H_{EASY}), ed uno avente direzione orientata coincidente con l'hard-axis (vedi H_{HARD}). Tutto ciò è fondamentale per portare a zero la barriera energetica, quindi per magnetizzare opportunamente lo strato libero del MTJ.

In particolare, è possibile notare come in corrispondenza di campi magnetici esterni nulli ($H_{EASY} = H_{HARD} = 0$), sia la barriera energetica Eb, sia la magnetizzazione corrente della cella (Hsw), sono massime. Inoltre, la presenza di uno solo di questi due campi comporta una riduzione solo parziale della barriera energetica di cui sopra, nonché la loro singola presenza non è utile alla realizzazione dello switching del stato di bit.

In realtà, la definizione di un campo magnetico di hard-axis permette di ridurre la barriera Eb in modo tale che il successivo scorrimento all'interno della Bitline di una corrente ridotta e generante un "campo magnetico di easy-axis diminuito", consenta di magnetizzare opportunamente lo strato libero della giunzione.

Pertanto, è evidente che in assenza di un campo di hard-axis, la corrente che scorre nella Bitline genera un campo di easy-axis energeticamente insufficiente a magnetizzare in modo opportuno il dispositivo MTJ. Tutto ciò è riassunto nella Fig. 2.4 di sotto.

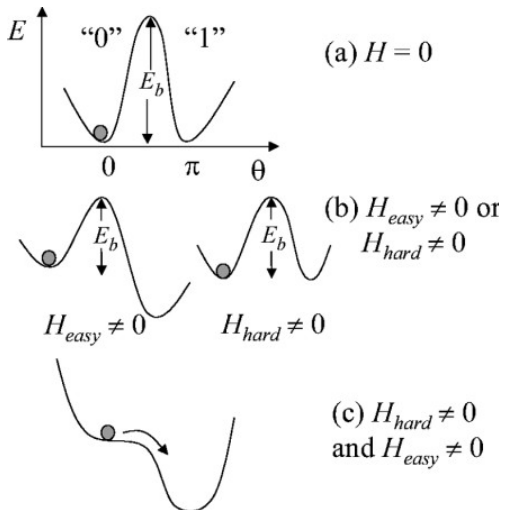


Figura 2.4. [1] Nella parte superiore, la assenza di campi magnetici esterni e la conseguente barriera energetica Eb, consente alla cella di conservare la propria magnetizzazione. Nella parte centrale è mostrato come la presenza di uno solo dei campi magnetici esterni non sia sufficiente a modificare la magnetizzazione del FreeLayer del MTJ (H_{EASY} comporterà solo un "aumento di energia potenziale" tra lo stato di partenza e quello finale di commutazione, che però risulterà inutile se tra questi esiste una barriera energetica; H_{HARD} permetterà di abbattere Eb, ma ciò è del tutto inefficace se lo stato iniziale e finale sono energeticamente uguali). Infine, nella parte bassa è rappresentata la situazione in cui sono presenti entrambi i campi magnetici esterni.

In realtà, questo fenomeno è adoperato anche per realizzare la selezione di una singola cella di memoria, e cioè: il campo necessario per le operazioni di switching risulta ridotto per tutte le celle collegate alla Digitline percorsa da corrente.

Tuttavia, risulta effettivamente programmabile solo la cella che si trova in corrispondenza dell'intersezione della stessa Digitline presa in considerazione, e della Bitline percorsa da corrente. Tutte le celle per le quali si rivela attivata solamente ed in modo totalmente esclusivo la Bitline o la Digitline, sono dette half-selected bits e non sono programmabili.

La presenza di un'unica barriera energetica per il Free Layer, è una situazione del tutto ideale poiché a causa di difetti di fabbricazione è possibile che ve ne siano più di una. Ciò implica che in ogni singola cella,

lo “strato programmabile” possa essere magnetizzato stabilmente anche secondo direzioni orientate che rappresentano però valori logici spuri.
Le variazioni da una cella all’altra dei campi magnetici richiesti per la commutazione, influenzano l’abilità di produzione di memorie con architetture matriciale ed ad alta densità di integrazione (vedi Fig. 2.5).

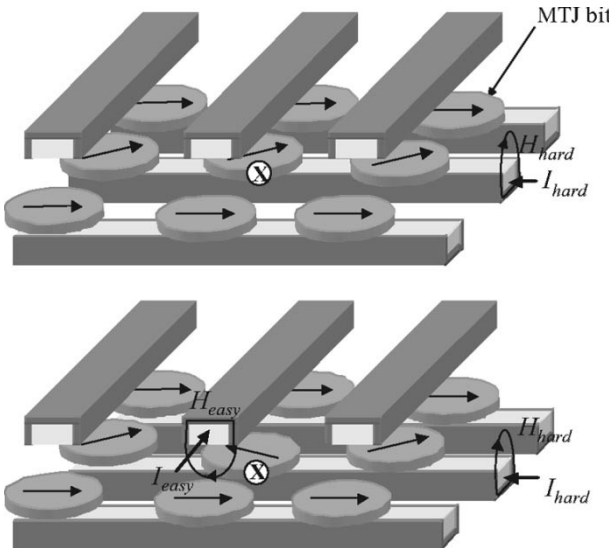


Figura 2.5. [1] Architettura Cross-Point di una MRAM. Un singolo bit è definito in corrispondenza dell’intersezione delle linee di Bitline e Digitline. La corrente di h-ax. (I-hard) genera il campo magnetico di h-ax. (H-hard), così da ridurre la barriera di energia. La corrente di e-ax.(I-easy) genera il campo magnetico di e-ax. (H-easy), che permette di azzerare completamente la Eb, quindi di magnetizzare opportunamente il FreeLayer del MTJ selezionato.

Durante un’operazione di scrittura, tutte le celle che appartengono alla linea di Digitline, e tutte quelle disposte sulla linea di Bitline, attivate per la selezione di un singolo MTJ generico, sono esposte ad uno dei due campi magnetici esterni utilizzati per lo switching. Tale campo (quindi quello di easy-axis, o di hard-axis) dovrà avere intensità minore del valore minimo utile per lo switching dello stato di bit, perché altrimenti le stesse celle di cui sopra potrebbero subire disturbi della loro magnetizzazione. Ovviamente, l’intensità del campo magnetico (di “half-selective”, se così si può dire) dovrà avere un valore che non potrà andare al di sotto di quello

minimo richiesto per attuare lo switching della cella effettivamente selezionata.

Il discorso precedente è sinteticamente rappresentato dalla Fig. 2.6, la quale mostra la “Regione Operativa della Corrente”. In questa immagine è possibile notare la curva matematica (noto come “astroide” [32]) che definisce, in funzione delle intensità dei campi di h-ax. e e-ax., quali sono le combinazioni di valori limite degli stessi, per cui in corrispondenza di valori superiori a quelli di soglia, si ha la scrittura della cella, mentre per valori inferiori, non risulta alterata la magnetizzazione del “Free Layer” del MTJ.

A causa della non perfetta precisione del processo di produzione dei dispositivi, ed a causa dei piccoli difetti nei materiali, ad un insieme di celle di MRAM è possibile associare una “distribuzione di campi magnetici di commutazione”, la quale sarà caratterizzata da un certo valore di deviazione standard σ (al tal proposito, i rettangolini in grigio presenti in corrispondenza degli assi di coordinate, rappresentano esempi di come per celle dello stesso dispositivo possano esistere differenti combinazioni di campi di switching di h-ax. e di e-ax, i cui valori d’intensità definiscono delle distribuzioni rappresentabili proprio come aree).

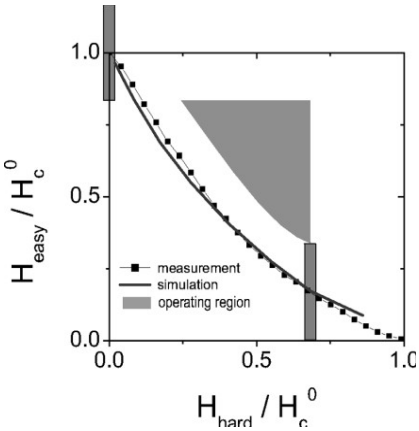


Figura 2.6. [1] Astroide dei campi magnetici utili per lo switching dello stato di una cella di MRAM

(vedi curva di simulazione e punti rappresentanti misurazioni reali in regime "Quasi-Stazionario").

A questo punto, tenendo conto che valori eccessivamente elevati dell'intensità dei campi applicati potrebbero disturbare la magnetizzazione delle celle half-selected, è possibile parlare di regione di funzionamento per i campi esterni (è quella in grigio al di sopra dell'astroide), intendendo per questa l'insieme delle combinazioni di valori degli stessi campi per cui le celle risultano programmate correttamente, senza la generazione di errori in quelle con selezione a metà.

L'esistenza di una distribuzione per i campi di switching è dovuta a numerosi fattori, come la forma, la dimensione ed il materiale della cella; o ancora, il processo di fabbricazione della stessa [33-36]. Si dimostra, che ottimizzando questo insieme di fattori, l'anisotropia di forma delle celle realizza una sola ed unica barriera energetica [37], cosicché la programmazione avviene utilizzando gli stessi identici campi esterni per tutte le celle della memoria (in questo caso, la qualità del dispositivo di memoria è elevata).

Sempre a proposito della distribuzione di switching, un'altra causa di questo fenomeno è rappresentata dall'interazione magnetica tra celle vicine [38]: i campi magnetostatici derivanti dalle magnetizzazioni delle celle vicine a quella effettivamente da programmare, possono sommarsi/sottrarsi a quelli esterni applicati in fase di scrittura (in realtà questo disturbo influenza molto poco la densità d'integrazione delle MRAM [38]).

La comprensione degli aspetti di micro-magnetismo che caratterizzano la giunzione a tunnel magnetoresistivo, è fondamentale sia per la riduzione dell'estensione della distribuzione dei campi magnetici di switching, sia per migliorare l'effetto prodotto dal campo esterno di h-ax.

Un modello per simulare il comportamento micro-magnetico degli MTJ di una MRAM è stato sviluppato, e prevede che le stesse celle di memoria siano di forma ellittica con estensione d'area di 0,6 X 1,2 μm^2, nonché abbiano il proprio Free Layer con spessore pari a 4 nm, realizzato in una lega di Nichel, Ferro e Cobalto. Infine, il modello prevede anche che le linee di Bitline e di Digitline abbiano spessore pari a 0,9 µm.

Il modello di cui sopra è utilizzato per simulare il processo di inversione della direzione di magnetizzazione che si sviluppa durante la fase di scrittura. In particolare, come mostrato nella Fig. 2.7 (parte sinistra), gli impulsi di corrente utilizzati per lo switching dello stato di bit della cella, richiedono un tempo esattamente uguale a 2 ns per raggiungere il loro valore finale (si assume cioè che al tempo t = 0 ns, entrambe le correnti di Bitline e Digitline abbiano valore pari a quello richiesto per eseguire l'operazione di scrittura della cella, e che la generazione degli stessi impulsi sia avvenuta all'istante t = -2 ns).

La Fig. 2.7 (parte destra) mostra che il processo d'inversione della direzione di magnetizzazione del Free Layer si completa in un tempo pari a 2 ns (inizia all'istante t = 0 ns, termina per t = 2 ns), nonché l'inversione si sviluppa dapprima nella parte centrale della cella, quindi si diffonde fino a coprire tutta l'area ellittica della cella di memoria.

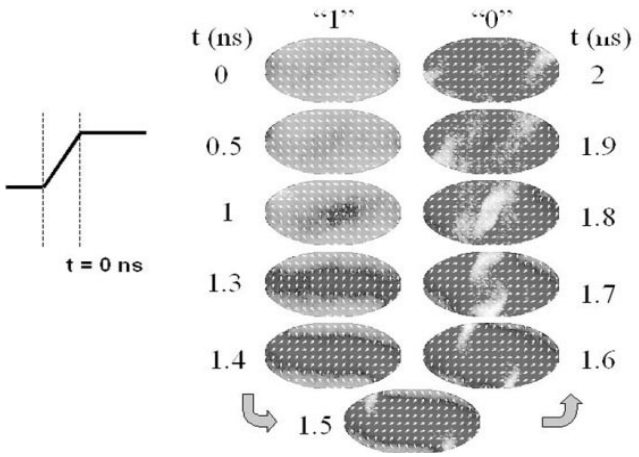


Figura 2.7. [1] Nella parte sinistra è mostrata la transizione [valore iniziale => valore finale] compiuta dalle correnti di Bitline e Digitline: essa avviene in un intervallo di tempo pari a 2 ns, e termina all'istante t = 0 ns. Nella parte destra è invece rappresentato il processo di inversione della direzione di magnetizzazione del FreeLayer della cella di MRAM (anch'esso si completa in un intervallo di tempo pari a 2 ns, nonché termina in corrispondenza dell'istante t = 2ns).

Siccome le proprietà mostrate dalle celle nei processi di switching che avvengono con alte velocità sono molto importanti per lo sviluppo di MRAM ad alte frequenze di funzionamento, allora si è pensato di testare celle a MTJ con area 0,45 X 1,35 μm^2, per il caso di processi di switching che avvengono in intervalli temporali misurabili con una scala in nanosecondi.

Ovvero, le celle di memoria di cui sopra sono state testate utilizzando impulsi di corrente di easy/hard-axis con durata pari a circa 20 ns (di conseguenza il processo di switching avviene in un identico intervallo temporale), ottenendo così l'astroide (dei valori medi delle correnti necessarie allo switching) rappresentato in Fig. 2.8 (vedi curva passante per i punti pieni).

In realtà, in Fig. 2.8 è mostrato anche un altro astroide (vedi curva passante per i punti vuoti), ottenuto utilizzando un campo magnetico di hard-axis costante, generato appunto introducendo all'interno della MRAM un magnete esterno.

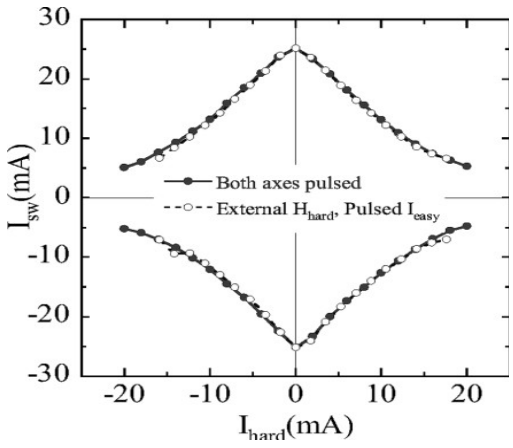


Figura 2.8. [1] Astroide dei valori medi delle correnti necessarie allo switching (vedi curva passante per i punti pieni) ed astroide delle correnti necessarie allo switching in caso di campo esterno di hard-axis costante (vedi curva passante per punti vuoti).

Da queste simulazioni, emerge che i processi di switching sviluppati in intervalli temporali di poche decine di nanosecondi [39], [40], non producono effetti giromagnetici che potrebbero alterare lo spin degli elettroni presenti nei layer ferromagnetici MTJ, pertanto le MRAM offrono elevate velocità di scrittura/programmazione delle proprie celle di memoria.

L'unica obiezione che si può formulare in questi test, è relativa al fatto che sia la Bitline, sia la Digitline, sono state implementate come linee con spessore/larghezza piuttosto elevata, nonché queste caratteristiche le rende incompatibili con le dimensioni ridotte dei dispositivi reali MRAM.

La minimizzazione dello spessore delle linee utilizzate per la programmazione deve pertanto coincidere con il rivestimento delle stesse con un strato di materiale ferromagnetico, detto "cladding layer" (c.l.).

Il compito del "c.l." è fondamentale poiché reindirizza e concentra il flusso di campo magnetico prodotto dal passaggio di corrente nel conduttore, dalla sua parte più esterna verso la parte più vicina alla cella ad MTJ: questo accorgimento consente all'incirca di dimezzare l'intensità delle

correnti di switching richieste per generare i campi magnetici esterni di easy/hard-axis.

Il discorso precedente è sintetizzato nell'astroide di Fig. 2.9, nella quale sono messe a confronto una coppia di linee senza il c. l., con una coppia avente il rivestimento in questione. La curva inferiore mostra che: dato uno stesso valore di corrente di h-ax, è possibile programmare una cella di MRAM con un valore di corrente di e-ax. inferiore in presenza di c. l., rispetto a quello effettivamente richiesto nel caso in cui il c. l. invece mancasse.

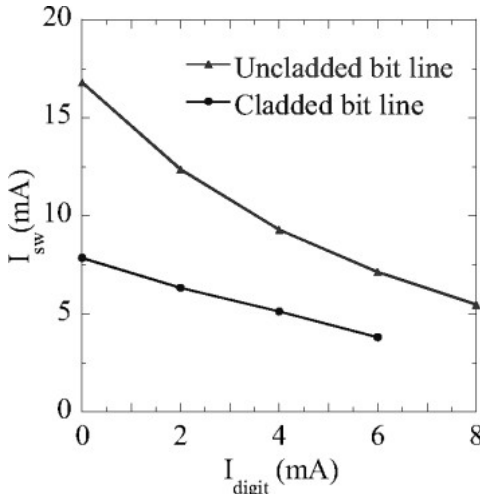


Figura 2.9. [1] Astroide delle correnti di switching per il caso di Digitline e Bitline "cladded" ed "uncladded". Si noti come, utilizzando linee "schermate", è possibile rilevare una diminuzione della corrente di easy-axis necessaria alla programmazione (vedi curva inferiore).

Inoltre, la presenza del cladding layer sulle linee di programmazione consente di evitare che campi magnetici prodotti da altre connessioni circuitali del dispositivo di memoria, possano interferire con i processi di switching in esecuzione.

La produzione di MRAM nella forma di dispositivi sempre più piccoli comporta alcuni problemi, e cioè: se da un lato la riduzione delle dimensioni degli elementi circuitali interni determina un aumento delle

intensità dei campi magnetici (l'intensità di questi stessi aumenta infatti al diminuire dello spessore della linea che lo genera), dall'altra l'utilizzo di connessioni più sottili ha come diretta conseguenza l'aumento delle intensità di correnti necessarie allo switching. Questi due fattori a loro volta determinano un più alto consumo di potenza.

In questo contesto inoltre, bisogna tener presente che la riduzione delle dimensioni della singola cella, ha come suo effetto l'aumento della ripidità dell'astroide associato al dispositivo MRAM, indicante appunto che solo le celle che avranno un alto valore di intensità per la corrente di e-ax., in corrispondenza ovviamente di un opportuno basso valore di corrente di h-ax., saranno effettivamente programmate. Le altre invece no (esiste quindi una più netta separazione tra la cella selezionata e quelle half-selected).

In definitiva, i progressi riguardanti la realizzazione delle celle e le loro connessioni, garantiscono un discreto margine di affidabilità dei processi di scrittura delle celle ad MTJ.

Tuttavia, la riduzione delle dimensioni di cella determina un ulteriore problema, quale quello della stabilità dello stato di bit in presenza di fluttuazioni termiche dovute all'agitazione (termica) delle molecole.

E cioè, al crescere della temperatura, un generico sistema devia casualmente dalla sua condizione di equilibrio (effettiva definizione di fluttuazione termica) [9], e questo fenomeno comporta l'abbattimento completo e non parziale per le celle half-selected di una MRAM, della barriera energetica Eb (si tenga presente che le fluttuazioni termiche sono il risultato di disturbi). Tutto ciò, induce a considerare come concreta l'ipotesi che lo switching dello stato di bit possa avvenire anche per celle non effettivamente selezionate.

Infine, è evidente che l'ottimizzazione degli aspetti relativi alle dimensioni delle celle, ai campi di switching ed alla distribuzione a loro associata,

determinano più o meno rilevanti problematiche legate agli effetti termici di cui sopra.

2.4 – Architettura e Integrazione delle MRAM.

Nella Fig. 2.10 è mostrata la sezione trasversale di una cella integrata di MRAM, che utilizza architettura 1T1MTJ. Si nota subito che il modulo di MRAM è inserito tra gli ultimi due strati di metallo [41]: per questo motivo si parla di modulo di Back-end, e cioè di un modulo che è integrato nel sistema complessivo solo quando tutta la circuitazione CMOS, indicata come modulo CMOS di Front-End, è stata completamente implementata.
Siccome quindi il processo di fabbricazione della componente di front-end non è alterata, questo approccio di strutturazione a due blocchi è ideale per la creazione di sistemi embedded, per i quali è sovente richiesto che il blocco di memoria sia incluso già all'interno dello stesso sistema.
Inoltre, un ulteriore punto di merito di questo approccio è rappresentato dal fatto che il processo di realizzazione della circuitazione CMOS è ben separato da quello con cui sono implementati i componenti in materiale magnetico ad alta specificità degli MTJ.

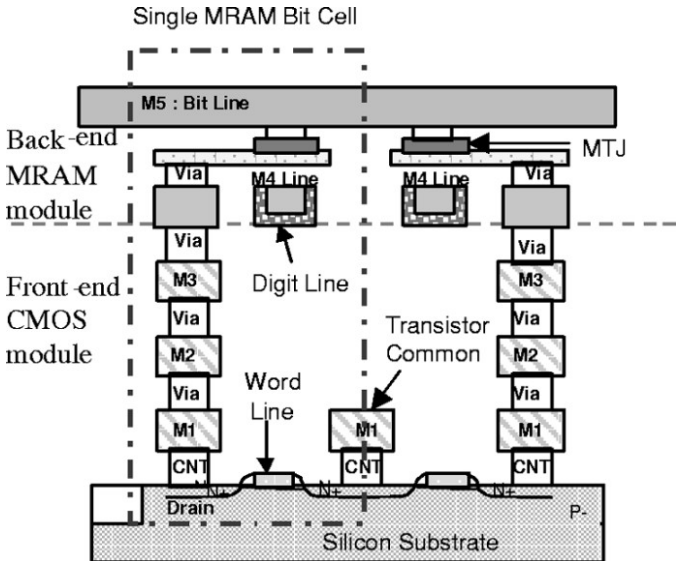


Figura 2.10. [1] Sezione trasversale di due celle per un dispositivo MRAM. La singola cella è inserita in corrispondenza degli ultimi due strati dello stack visibile in figura, cosicché si parla di struttura con Blocco di Back-End (componente di memoria del sistema) e Blocco di Front-End (circuitazione generale CMOS).

La circuitazione CMOS è necessaria per realizzare le operazioni di scrittura e lettura delle celle di MRAM. In particolare, essa è fondamentale per l'opportuna definizione dei flussi di corrente sulle numerose connessioni di Bitline e Digitline, i cui impulsi appunto consentono la programmazione contemporanea ed in un unico ciclo di più celle (sono salvate word da 16 o più bit).

Le operazioni di lettura invece, sono realizzate accendendo i transistor di isolamento attraverso le linee in silicio policristallino di Wordlines (nota sono disposte parallelamente alle Digitlines), che appunto ne controllano i gate.

In particolare, la lettura dello stato di bit di una singola cella prevede che siano attivate contemporaneamente le corrispondenti Bitline e Wordline: una word di 16 o più bit è disponibile in uscita al dispositivo se si attivano

altrettante Bitlines, quindi la Wordline condivisa tra le celle in considerazione.

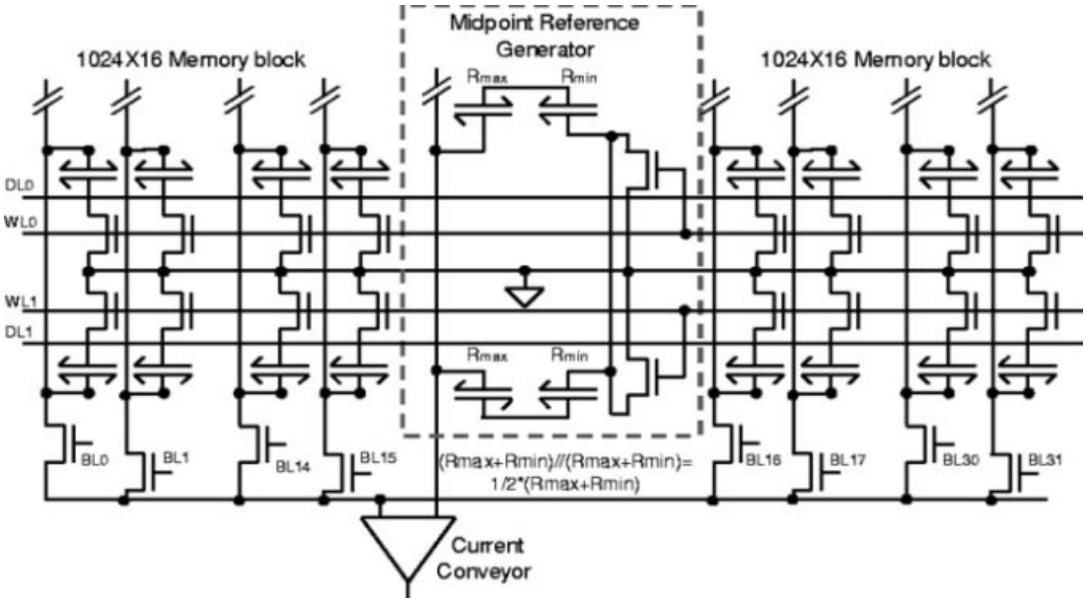


Figura 2.11. [1] Core di una MRAM con architettura 1T1MTJ, in cui il blocco di memoria è affiancato da quello di Generazione del Valore di Riferimento.

La linea di colonna sul quale è presente il valore della resistenza di riferimento, consente di realizzare il confronto tra questo stesso e quello posseduto da ciascuna cella disposta sulla medesima riga attiva (cioè quelle che condividono la Wordline a valore alto), cosicché è possibile determinare lo stato (resistivo) di ogni singolo bit memorizzato.

Il blocco che genera il valore resistivo di riferimento è costituito da quattro celle ad MTJ: due di questi hanno allineamento parallelo, ed altri due hanno allineamento antiparallelo. Pertanto, il valore di riferimento che risulta da questa disposizione è pari esattamente al valore medio delle resistenze di R_{MAX} e R_{MIN} definite dagli MTJ. Si noti infatti, come all'interno del dispositivo generatore sono disposti in parallelo due rami-serie, ciascuno dei quali è a sua volta costituito da una resistenza massima (MTJ

con allineam. antiparrallelo) in serie con una minima (MTJ con allineam. parallelo).

A questo punto il circuito di lettura determina il valore logico alto/basso del singolo bit a seconda appunto che il suo valore di resistenza sia rispettivamente maggiore/minore del valore di riferimento.

Disporre ogni blocco di celle strettamente vicino ad un “proprio” circuito di generazione del valore (di resistenza) di riferimento, come fatto nell’implementazione di Fig. 2.10, fa sì che ciascuno di questi moduli di celle abbia un “proprio tipico” valore di riferimento, dato appunto che quest’ultimo subirà le stesse alterazioni dovuti a fattori termici ed ad altri elementi esterni che interessano le celle del blocco in esame.

L’esecuzione di un’operazione di lettura di una singola cella può quindi essere sintetizzata nei seguenti passi:

1- la selezione avviene portando a valore alto sia la Bitline (è una linea di colonna per la struttura a matrice), sia la Wordline della cella in esame;

2- tenendo sempre presente che ogni blocco di celle ha un proprio circuito di generazione del valore di riferimento di resistenza, allora si attiva anche la linea di output del circuito di riferimento;

3- infine, il circuito di lettura rileva le correnti presenti sulla Bitline e su quella di Riferimento, e produce così in uscita un segnale in tensione che rappresenta il valore logico basso/alto appena letto.

Si noti, come le operazioni di lettura non sono distruttive, quindi le MRAM non prevedono l'esecuzione di operazioni di ripristino dei valori logici di cella.

Capitolo 3 – Memorie Ferroelettriche.

Precisazioni.

Il seguente capitolo è stato scritto riassumendo e traducendo in parte l'articolo [1], dal quale inoltre sono state tratte anche le figure ed i grafici che ivi compaiono.

Le memorie FeRAM (Ferroelectric RAM) presentano celle ciascuna costituita da un "capacitore ferroelettrico" e da un transistor/una coppia di transistor, che consentono di accedere alla cella e leggerne il valore logico salvato.
In particolare, la fase di lettura avviene utilizzando un sense amplifier (s.a.) che realizza un confronto tra la tensione presente sulla "linea di colonna" cui è connesso il capacitore ferroelettrico, ed un voltaggio di riferimento.
Come già detto in precedenza, la tecnologia che ha avuto maggiormente successo nell'ambito delle memorie non volatili è quella dei dispositivi di memoria a floating-gate (F-G).
Tuttavia, l'interesse sviluppato verso le FeRAM è giustificato da due aspetti principali, ovvero: (1) il tempo di programmazione delle celle è piuttosto ridotto; (2) il consumo di potenza di questa categoria di memorie è molto basso.

Nonvolatile Memory	Area/Cell (normalized)	Read Access-Time	Write (prog.) Access-Time	Energy* per 32b Write	Energy* per 32b Read
EEPROM	2	50ns	10μs	1μJ	150pJ
Flash Memory	1	50ns	100ns	2μJ	150pJ
Ferroelectric Memory	5 (†)	100ns	100ns	1nJ	1nJ

Tabella 3.1. [1] Confronto tra tecnologie implementative di memorie non volatili. (*) La tabella mostra anche l'energia richiesta per le fasi di scrittura e di lettura di un dispositivo con capacità di memoria pari a 32-bit. E' importante precisare che questi valori potrebbero differire se le celle di memoria fossero direttamente integrate all'interno di circuiti più complessi. (†) Il valore assunto dal parametro Area-per-Cella per le memorie FeRAM può raggiungere quelli tipici delle EEPROM, nel caso in cui le singole celle di FeRAM utilizzassero processi di fabbricazione più sofisticati quanto più convenienti, come appunto l'utilizzo di "capacitori ferroelettrici a stack" (vedi esempio Fig. 3.2).

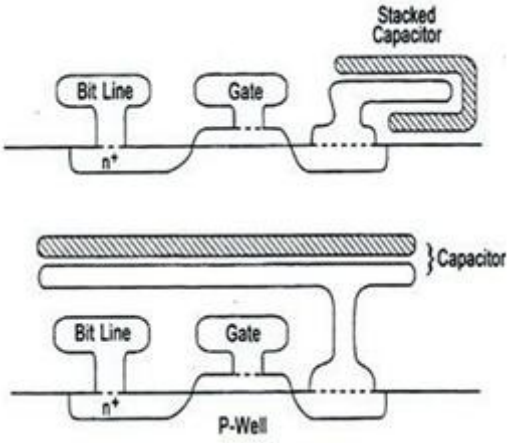


Figura 3.2. [2] Confronto tra una cella di memoria DRAM avente il proprio transistore di accesso collegato ad un capacitore con struttura a stack (immagine superiore) ed una cella di DRAM avente il proprio transistore d'accesso collegato ad un semplice condensatore piano (immagine inferiore). E' evidente la riduzione di spazio conseguente all'adozione di una struttura a stack per il condensatore di memorizzazione del bit.

La Tab. 3.1 mostra il confronto realizzato tra componenti di memoria FeRAM, EEPROM e Flash. Quest'ultimi due dispositivi di memoria mostrano stesse performance per quanto riguarda l'energia ed i tempi richiesti per la fase di lettura, mentre presentano valori differenti per

quanto riguarda la densità d'integrazione, e l'energia/tempi richiesti per la programmazione delle loro celle.

Le tecnologie EEPROM e Flash sono sicuramente più vantaggiose rispetto alle FeRAM, se si considerano contesti applicativi che richiedono nelle fasi di lettura un basso consumo di potenza e veloci tempi di reperimento dati. In aggiunta, sempre rispetto ai dispositivi FeRAM, essi offrono una maggiore densità d'integrazione [6, 7].

Le memorie FeRAM, al contrario rappresentano una migliore soluzione per contesti applicativi che richiedono veloci tempi di programmazione delle celle, ed un complessivo basso consumo di potenza (e cioè non solo in fase di lettura, come garantito invece dalle memorie a F-G di cui sopra).

Inoltre, le FeRAM sono maggiormente adatte ad essere integrate all'interno di circuiti complessi, vedi per esempio l'implementazione di SoC (System-On-Chip) [8, 9] utilizzati dalle applicazioni embedded.

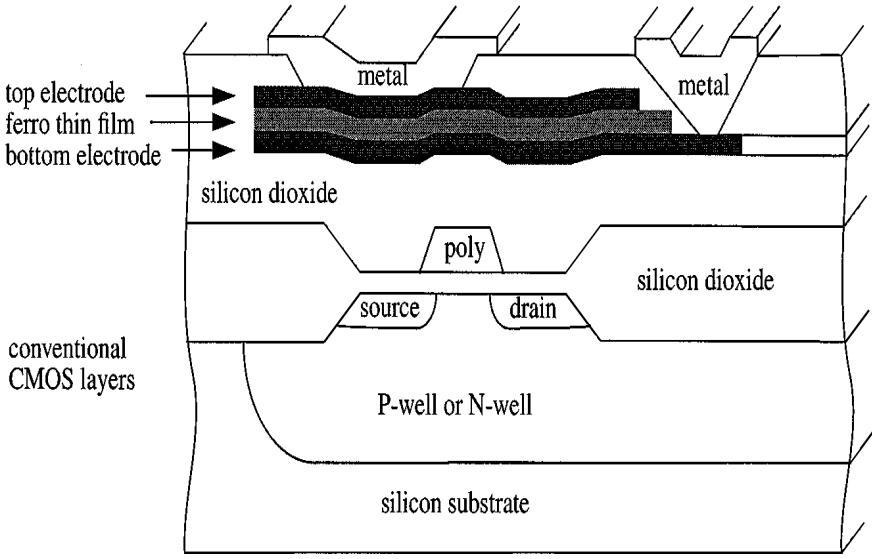


Figura 3.3. [1] Gli strati che compongono il condensatore ferroelettrico sono disposti al di sopra di un convenzionale transistor (n/p)-MOS.

Il capacitore ferroelettrico tipico di una cella di FeRAM è implementato utilizzando un opportuno insieme di layers (due elettrodi separati da una

sottile pellicola in materiale ferroelettrico), disposti al di sopra degli strati realizzati con processi CMOS (vedi Fig. 3.3).

In questo modo quindi, nello stesso chip coesistono circuiti utilizzanti segnali digitali (vedi appunto quelli CMOS), con quelli a segnali analogici (nota: le parti che non utilizzano i capacitori ferroelettrici sono opportunamente isolate dai medesimi).

I principi di funzionamento alla base delle operazioni eseguibili sui condensatori ferroelettrici sono fondamentalmente gli stessi adoperati dalle memorie ferromagnetiche, note anche come memorie a “nuclei ferromagnetici” (vedi Fig. 3.4).

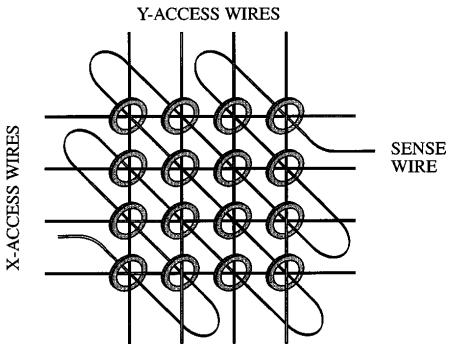


Figura 3.4. [1,13] Struttura matriciale di una memoria ferromagnetica.

In realtà, questa tecnologia implementativa è stata velocemente abbandonata in virtù di tre principali ragioni, e cioè: l’utilizzo di toroidi di ferrite [3] per l'implementazione dei nuclei ferromagnetici delle celle, determina la realizzazione di memorie eccessivamente voluminose, costose e con elevato consumo di potenza rispetto a quelle con tecnologia a semiconduttore ed ai dispositivi a tecnologia ferroelettrica.

In particolare, uno dei problemi più importanti che affliggono le memorie ferroelettriche riguarda la non perfetta coesistenza ed integrazione tra il

capacitore ferroelettrico e gli strati in materiale semiconduttore realizzati con processo CMOS. Questo aspetto ha conseguenze negative sulla densità d'integrazione delle memorie ferroelettriche, che appunto risulta inferiore rispetto a quella invece garantita dalle DRAM e dalle EEPROM.

3.1 – I capacitori ferroelettrici.

Per "capacitore ferroelettrico" si intende un condensatore che presenta uno strato di materiale ferroelettrico in luogo del comune layer di materiale dielettrico che separa le armature [14].

Questa caratteristica è di fondamentale importanza se si considera che i materiali dielettrici e quelli ferroelettrici si comportano in modo completamente differente nel caso di applicazione di un campo elettrico esterno. Ovvero, per un materiale dielettrico l'applicazione di un campo esterno, implica che le cariche positive e negative in esso presenti assumano delle posizioni diverse da quelle iniziali (così da formare un dipolo elettrico macroscopico). Esse però ritornano nella loro configurazione iniziale non appena il campo esterno cessa d'esistere.

Al contrario, per il caso di un materiale ferroelettrico, esiste già una polarizzazione elettrica spontanea (la quale dipende dalla struttura del cristallo del materiale), nonché applicando un opportuno campo elettrico esterno tale polarizzazione può essere invertita o opportunamente direzionata, senza che poi scompaia non appena vi sia assenza di campo esterno.

I "materiali perovskitici", cioè ossidi di sintesi aventi struttura ABO_3 (con A = Metallo Alcalino Terroso, B = Metallo di Transizione e O = Ossigeno) [4] sono quelli più comunemente utilizzati per le loro proprietà ferroelettriche.

Nella Fig. 3.5 è mostrata la struttura del perovskite $Pb(Zr_xTi_{1-x})O_3$, caratterizzato per avere al centro del reticolo alternativamente un atomo di Zirconio (Zr) o di Titanio (Ti).

La figura mostra come l'applicazione di un campo elettrico esterno sia utile a spostare l'atomo centrale del cristallo in una nuova posizione stabile, direzionando conseguentemente nel modo desiderato lo stato di polarizzazione dello stesso cristallo.

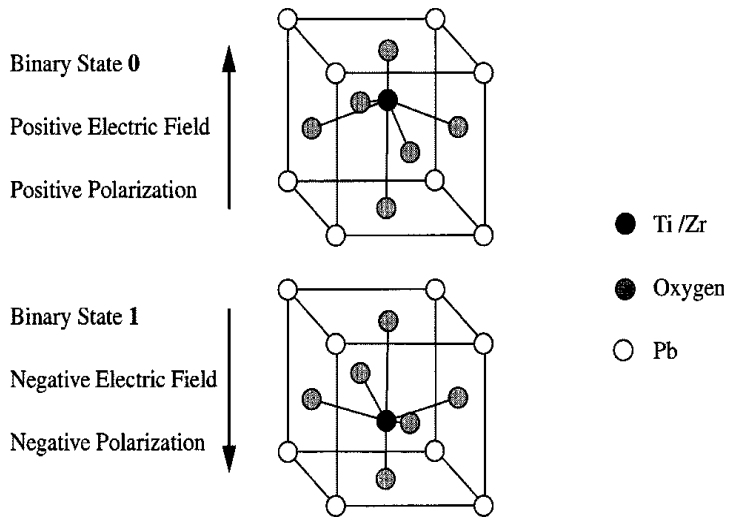


Figura 3.5. [1] Stati stabili di polarizzazione elettrica per un cristallo di materiale ferroelettrico. L'applicazione di un campo elettrico esterno consente di indirizzare opportunamente la polarizzazione spontanea che caratterizza i cristalli del materiale.

Sebbene il dipolo elettrico associato ad un singolo cristallo sia piuttosto piccolo, la polarizzazione netta derivante dalla somma vettoriale dei dipoli di un insieme di cristalli allineati, è sufficientemente grande da poter essere rilevata mediante un'architettura circuitale con sense amplifier.

Ad ogni modo, la conseguenza più importante della polarizzazione netta del materiale ferroelettrico è che il suo dipolo totale direzionato opportunamente, determina una carica non nulla per unità di area nel

capacitore ferroelettrico all'interno del quale lo stesso strato in esame risulta essere inserito. Siccome la polarizzazione del materiale ferroelettrico è presente e stabile anche in assenza di campo esterno, allora lo è anche la carica totale del condensatore.

Sempre rimanendo nell'ambito dei capacitori ferroelettrici, la carica presente sulle loro armature dipende strettamente dal voltaggio applicato ai capi del materiale ferroelettrico durante la fase di programmazione.

Nella Fig. 3.6 è mostrato il loop d'isteresi di un generico capacitore ferroelettrico, cioè il grafico rappresentante la dipendenza della carica presente sulle armature del condensatore (asse delle ordinate) in funzione della tensione applicata ai capi dello stesso (asse delle ascisse).

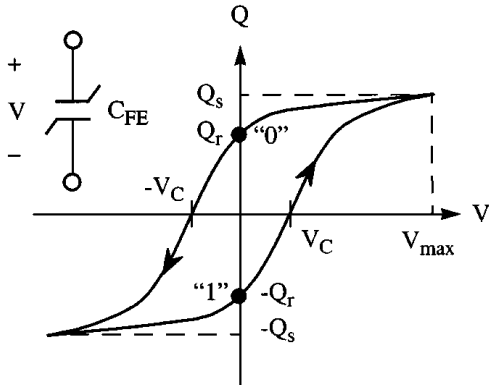


Figura 3.6. [1] Loop di isteresi e simbolo circuitale (in alto a sinistra) di un capacitore ferroelettrico. Nella figura, ±Vc rappresentano i valori di tensione necessari da applicare ai capi dei singoli capacitori ferroelettrici per poter orientare opportunamente la polarizzazione del materiale ferroelettrico, ed indurre così una carica ±Qr che rappresenti coerentemente lo stato logico "0" o "1" rispettivamente. I valori di tensione indicati con ±Vmax sono invece quelli che consentono di indurre una carica massima di saturazione ±Qs. Infine, con ±Qr sono indicati i valori della carica "di rimanenza".

Dalla figura si capisce chiaramente che la carica del capacitore (e cioè quella presente sulla sua armatura superiore) assume stabilmente il valore (+Qr) o (-Qr) (a seconda ovviamente della tensione esterna

precedentemente applicata) anche quando la tensione esterna è nulla, implicando dunque che il capacitore ferroelettrico permane nello stato di bit “0” od in quello “1” rispettivamente anche una volta annullato il campo elettrico esterno.

Il processo di transizione dallo stato logico “0” a quello “1” avviene pertanto applicando esternamente al condensatore un impulso negativo di tensione ai suoi capi, cosicché: la direzione orientata della polarizzazione del materiale ferroelettrico di separazione delle armature risulti invertita, e conseguentemente sulla sua armatura superiore sia presente una carica pari a (-Qr). E’ ovvio che il processo per realizzare la transizione inversa dallo stato di bit “1” a quello “0” avviene in modo esattamente speculare a quello descritto per il passaggio [stato logico 0 → stato logico 1] (si applica un impulso positivo di tensione ai capi del capacitore, così da determinare sulla armatura superiore la presenza di carica pari a (+Qr)).

Detto ciò, dal loop di isteresi in figura è possibile notare come le transizioni verso uno dei due stati stabili di bit (che si manifestano appunto in corrispondenza dei voltaggi di coercizione ±Vc) non siano particolarmente repentine e nette, e questo perché l’instaurazione del campo elettrico esterno consente di direzionare opportunamente solo una parte dei dipoli elettrici dei cristalli dello strato di separazione delle armature.

Questa peculiarità dei condensatori ferroelettrici impatta negativamente sulla robustezza degli stati di bit delle singole celle, e cioè: ipotizzando un’architettura matriciale per una FeRAM, siccome la polarizzazione dello strato di materiale ferroelettrico può essere alterata anche con valori di tensione esterna inferiori alla metà del valore Vmax (valore di tensione utile appunto a definire sulle armature del condensatore la carica massima Qs, detta di saturazione), allora è molto probabile che un’operazione di

switching di una qualsiasi cella, comporti l'alterazione degli stati logici delle celle che sono disposte sulla stessa riga e o sulla stessa colonna di quella da scrivere.

Per correggere questo problema è necessario modificare il processo di fabbricazione dei materiali ferroelettrici in modo tale che il ciclo di isteresi associato ai relativi condensatori, risulti avere forma prossima a quella di un quadrato (vedi loop di isteresi di un core ferromagnetico).

In realtà, come mostrato nello schema di Fig. 3.7, una prima e semplice soluzione al problema di cui sopra consiste nel disporre il capacitore ferroelettrico in serie con un transistore di accesso (nMos nel caso della figura). In questo modo, lo stato logico di cella potrà essere letto o scritto se e solo se il corrispondete transistore di accesso risulterà acceso, in caso contrario invece, il valore di bit salvato non verrà cambiato.

3.2 – Operazioni di scrittura e di lettura di una generica cella di FeRAM.

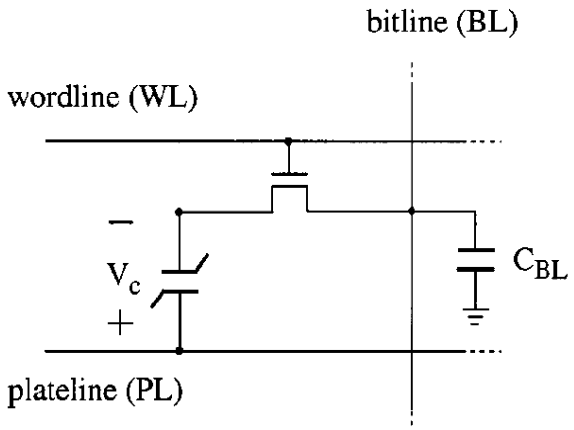


Figura 3.7. [1] Schema circuitale "1Transistor-1Capacitore Ferroelettrico" di una cella di FeRAM.

Come mostrato nella Fig. 3.7, il capacitore ferroelettrico di una cella di FeRAM presenta un'estremità connessa alla linea di Plateline (PL),

utilizzata per le operazioni di lettura/scrittura della singola cella, e l'altra collegata invece alla Bitline (BL) mediante un transistore d'accesso.

Le operazioni di scrittura dei valori logici “1” e “0” sono simili tra loro (vedi Fig. 3.8), e consistono: per il salvataggio di un valore logico alto, la tensione presente sulla BL è portata a V_{DD} (in realtà la tensione ai capi di C_{FE} sarà pari a $-V_{DD}$ in relazione ai segni indicati in figura 3.7) quindi viene generato un impulso di tensione di uguale ampiezza sulla linea PL che però non pregiudica la polarizzazione del condensatore ferroelettrico (durante la generazione dell'impulso, infatti, sulla BL è presente ancora una tensione pari a V_{DD}, quindi nell'intervallo di tempo per cui PL è valore alto, la caduta di tensione ai capi del C_{FE} è pari a zero, quindi la corrispondente polarizzazione non risulta modificata); nel caso in cui invece si voglia salvare un valore logico basso all'interno della cella, la linea di BL è mantenuta a tensione di 0V, mentre sulla linea di PL è generato un impulso di ampiezza pari a V_{DD} esattamente come nella situazione descritta precedentemente.

Ovviamente, per eseguire sia la prima che la seconda operazione, è necessario che il condensatore ferroelettrico non sia in stato di isolamento, e cioè sia raggiungibile dai livelli di tensione della BL grazie all’attivazione del transistore di accesso. Ciò implica, quindi, che le operazioni sono effettivamente realizzate se sulla Wordline che controlla il gate dell’nMos (vedi figura sopra) è presente un valore di tensione superiore a quello di soglia V_{TH} del transistore (in genere sulla WL sono presenti segnali di tensione circa pari a $V_{DD}+V_{TH}$, noti come “boosted V_{DD}” [15]).

E’ molto importante sottolineare come l’attivazione del transistor di accesso avvenga solo e soltanto quando la BL abbia raggiunto i valori di regime V_{DD} o 0V, a seconda appunto che si voglia scrivere rispettivamente

"0" o "1", e che il suo successivo spegnimento (riporta in isolamento il condensatore ferroelettrico) si verifica solamente quando la BL o la PL abbiano raggiunto realmente livelli di tensione pari a 0V.

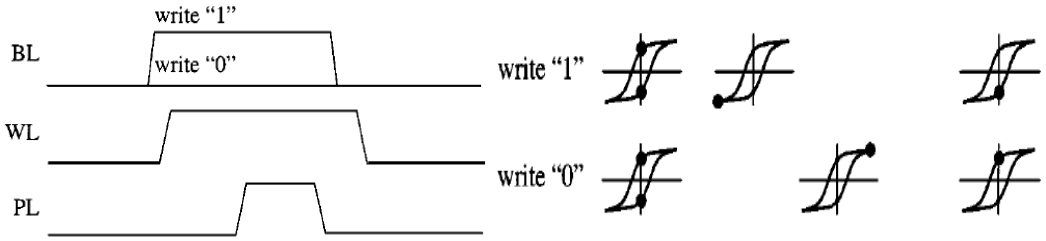


Figura 3.8. [1] Diagrammi temporali dei segnali presenti sulle linee BL, PL e WL durante le operazioni di scrittura dei valori logici "1" e "0" (parte sinistra [1]). Diagramma degli stati assunti dal capacitore ferroelettrico (parte destra [1]) durante le operazioni di scrittura di un valore logico alto (parte superiore) e di valore logico basso (parte inferiore). Lo stato finale del condensatore (cioè la carica di rimanenza dopo il processo di scrittura) non dipende dallo stato in cui si lo stesso si trova inizialmente.

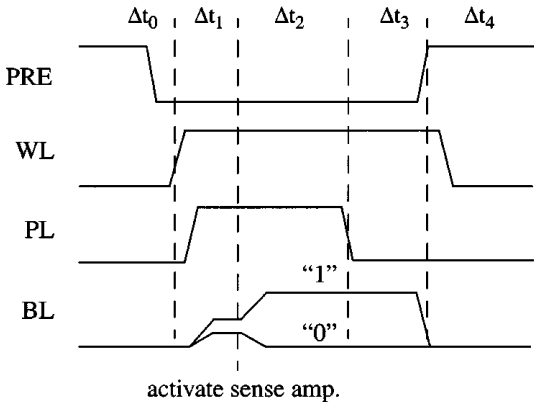


Figura 3.9. [1] Diagrammi temporali dei segnali presenti sulle linee di connessione della cella di FeRAM, durante la fase di lettura del valore logico.

Per quanto riguarda la fase di lettura, come indicato nella Fig. 3.9, essa comincia precaricando la BL al valore di tensione 0V, quindi si prosegue attivando la linea di WL, in modo da rimuovere il capacitore ferroelettrico dallo stato d'isolamento (vedi istante finale dell'intervallo Δt_0).

A questo punto, tenendo conto che la BL presenta una capacità parassita C_{BL}, allora la successiva attivazione della linea PL (vedi istante finale dell'intervallo Δt_1, in corrispondenza del quale PL è già a regime, cioè a V_{DD}), determina una suddivisione proporzionale di questa stessa tensione tra il capacitore ferroelettrico C_{FE} e quello C_{BL}, visto appunto che questa configurazione corrisponde proprio a quella di due condensatori disposti in serie (ed aventi quindi stessa carica), con a monte una tensione pari a V_{DD}. Tenendo conto che il capacitore ferroelettrico presenta un differente valore di capacità a seconda che il valore logico di cella sia "1" (C_1) o "0" (C_0), allora la tensione (V_X) presente sulla BL sarà pari a:

$$V_x = \begin{cases} V_0 = \dfrac{C_0}{C_0 + C_{\mathrm{BL}}} V_{\mathrm{DD}} \\ V_1 = \dfrac{C_1}{C_1 + C_{\mathrm{BL}}} V_{\mathrm{DD}} \end{cases}$$

L'operazione di lettura del valore logico prosegue con l'attivazione del sense amplifier, il quale porta la BL ad un livello di tensione pari a V_{DD} o a 0V, a seconda che il valore di tensione precedentemente rilevato dallo stesso sulla Bitline sia pari rispettivamente a V_1 o a V_0.

Infine, la fase di lettura termina con il ripristino del valore logico appena letto (vedi Fig. 3.9 gli intervallo Δt_3 e Δt_2 , fondamentali appunto per il ripristino rispettivamente dei valori logici "1" e "0").

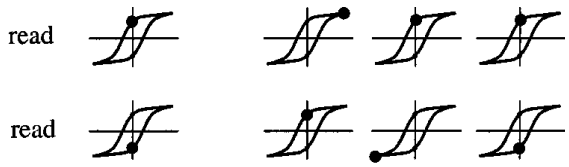


Figura 3.10. [1] Diagramma degli stati assunti dal capacitore ferroelettrico durante le operazioni di lettura di un valore logico basso (parte superiore) e di valore logico alto (parte inferiore).

3.3 – Generazione della tensione di riferimento.

I tre fondamentali compiti eseguiti dal circuito di s.a. sono: il rilevamento della tensione V_X presente sulla linea BL; il confronto dello stesso valore individuato con una tensione di riferimento V_{REF} (essa coincide idealmente con il valore medio dei valori V_1 e V_0 rilevabili sulla Bitline); ed infine, l'amplificazione della tensione presente sulla BL a valore di V_{DD} o a 0V a seconda che V_X sia rispettivamente maggiore o minore del valore di riferimento precedente.

Nella generazione del valore di riferimento si manifestano alcuni problemi, viste appunto le caratteristiche non ideali dei circuiti delle FeRAM.

A tal riguardo, è possibile in primo luogo notare come i valori di tensione V_X effettivamente rilevati dal s.a. differiscono da quelli teorici precedentemente calcolati, poiché il processo di fabbricazione delle celle "non è completamente uniforme", nonché nel tempo alcune caratteristiche circuitali tendono a rovinarsi (vedi ad esempio il fenomeno della "fatigue", indicante per l'appunto che il capacitore si degrada considerevolmente se lo stesso è sottoposto a numerosi cicli di lettura/scrittura [10]).

In seconda battuta, lo schema di generazione del voltaggio di riferimento deve esser tale da far fronte alle imperfezioni di fabbricazione dei capacitori ferroelettrici, che portano allo sviluppo di fenomeni come l'imprint [12] (cioè la tendenza di un condensatore ad assumere più facilmente una polarizzazione piuttosto che quella opposta, se per un lungo periodo di tempo non è stato alterato il proprio stato di bit, comportando così uno "shifting" lungo l'asse della Tensione, del loop di isteresi associato allo stesso C_{FE} [5]), o la relaxetion [11] (ovvero la perdita di carica di rimanenza in un intervallo di tempo dell'ordine di alcuni

microsecondi, se appunto il C_{FE} non è acceduto per più cicli consecutivi di lettura e/o scrittura. Ciò in genere comporta una diminuzione del valore reale di V_1, ed un aumento della tensione V_0).

Tutto ciò indica chiaramente che per un dispositivo di FeRAM non esiste in realtà un unico valore di V_{REF}, bensì più di uno a seconda delle variazioni di fabbricazione e del livello di degrado dei capacitori ferroelettrici.

3.4 – Primo esempio di circuito generatore della tensione di riferimento con schema ad un capacitore "sovradimensionato" per colonna (1C'/BL).

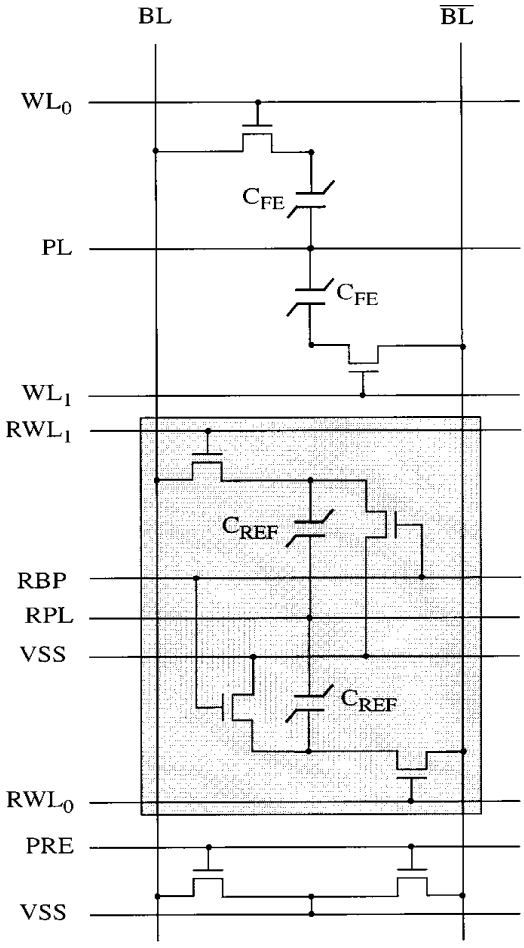


Figura 3.11. [1] Coppia di celle di riferimento (area colorata di grigio) per una colonna di FeRAM.

Il più comune schema circuitale di generazione del voltaggio di riferimento adoperato durante la fase di lettura delle celle che appartengono ad una stessa colonna, è rappresentato nella Fig. 3.11. In essa è visibile un'area grigia che mostra come il circuito in questione sia formato da ben due celle di riferimento, ciascuna delle quali è collegata in modo alternativo e mutuamente esclusivo alla linea di BL o $\overline{BL}$.

Inoltre, dal disegno si evince come ciascuna cella in questione possieda una propria Wordline, e cioè: i due capacitori ferroelettrici del circuito di riferimento (indicati come C_{REF}) sono accessibili attraverso le Wordline di Riferimento (o Reference Wordline) RWL_0 e RWL_1. La funzione di queste due linee è pertanto quella di rendere disponibile sulla BL o sulla linea complementare, il valore di riferimento a seconda appunto che la cella di memoria da leggere sia rispettivamente collegata alla seconda o alla prima delle due precedenti linee di colonna (ad esempio, la RWL_1 sarà adoperata per realizzare il confronto tra la V_{REF} e la tensione presente ai capi delle celle connesse alla linea di $\overline{BL}$, mentre la linea RWL_0 sarà adoperata durante la fase di lettura delle celle di memoria collegate alla linea di BL).

In questo modo dunque su una delle due linee di colonna è presente il valore logico della cella in lettura, mentre sulla quella complementare il valore di riferimento.

Le capacità dei condensatori presenti nel circuito di riferimento, sono più grandi di quelle che caratterizzano i capacitori delle normali celle di memoria. Questo consente di polarizzare ciascun condensatore di riferimento in modo tale da avere uno 0-logico, che in realtà corrisponde ad un valore tensione (appunto V_{REF}) circa pari al valore medio definito da V_1 e V_0 [15, 16].

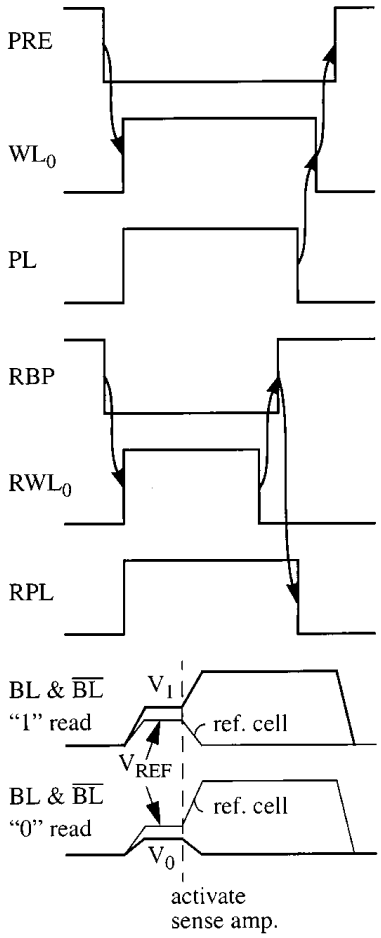


Figura 3.12. [1, 16] Diagrammi temporali dei segnali presenti sulle linee di connessione durante la fase di lettura di una generica cella appartenente alla Riga$_0$.

La prima fase del ciclo di lettura [16] prevede che sia le Bitline, che i nodi di memorizzazione delle celle di riferimento, siano precaricate con valori di tensione pari a 0V (nella parte iniziale dei diagrammi di sopra le linee di PRE e di RBP sono infatti entrambe a livello "1").

Ora, ipotizzando la lettura di una generica cella di memoria disposta sulla Riga$_0$, è necessario che siano portate al livello alto le connessioni WL_0 e RWL_0, e di conseguenza anche le linee PL e RPL (Reference-Plate Line, per la cella di riferimento appunto).

A questo punto, è quindi possibile disporre sulle linee di BL e di $\overline{BL}$ delle tensioni V_X e V_{REF}. In ultimo, l'attivazione del s.a. consente di portare a valore di V_{DD} o di 0V la linea di BL e di $\overline{BL}$, a seconda ovviamente dell'esito del confronto.

Per quanto riguarda i condensatori ferroelettrici di riferimento, l'operazione iniziale di precarica a 0V dei propri nodi di memorizzazione, fa sì che questi stessi capacitori sperimentino ai loro capi solo e soltanto variazioni positive di tensione, nonché la loro polarizzazione di rimanenza non risulta mai effettivamente alterata. Questo implica che tali condensatori lavorino in modalità non-switching e che conseguentemente soffrano solo in maniera del tutto marginale dei problemi di degrado descritti in precedenza.

In realtà, uno dei problemi che altresì affligge in maniera non trascurabile questo schema circuitale, è rappresentato dal fatto che il capacitore di riferimento, dopo l'attivazione del proprio segnale RPL, nell'ipotesi che il s.a. abbia portato la linea bitline a cui lo stesso condensatore è collegato a V_{DD}, risulterà avere ai propri capi un tensione pari esattamente 0V.

E cioè, quando RWL_0 ritorna a livello basso, il condensatore di riferimento presenta ad un capo (cioè il terminale collegato a RPL) una tensione pari a V_{DD}, mentre all'altro (ovvero il nodo di memoria connesso, mediante il transistore d'accesso, quindi alla BL o a $\overline{BL}$) una tensione che potrebbe essere fluttuante e non ben definita, ma molto prossima ad un valore di tensione alto.

Il fatto che ciascuna cella di riferimento possa disporre di un transistore aggiuntivo che resetti il valore logico del nodo di memorizzazione del condensatore di riferimento [16] prima che RPL sia disattivata, portando appunto la tensione presente sullo stesso a 0V, è di fondamentale

importanza (si noti come il “transistor di reset” sia attivato, portando appunto RBP a valore alto, subito dopo che la cella di riferimento sia stata isolata, cioè portando appunto RWL_0 a valore basso, e subito prima che RPL raggiunga un livello basso di tensione).

Questa soluzione implica ovviamente maggiore occupazione di spazio per l’implementazione del circuito di riferimento.

Per quanto riguarda il livello di tensione V_{REF}, è importante sottolineare come il valore di questo fondamentale parametro dipenda fortemente dalle dimensioni della capacità dei condensatori presenti nella cella di riferimento. Una stima teorica delle dimensioni capacitive di questi componenti è in genere molto difficile da realizzare, e tal proposito sono dunque necessari alcuni esperimenti per determinare in modo più preciso possibile il valore della capacità dei C_{REF}.

3.5 – Secondo esempio di circuito generatore della tensione di riferimento con schema a due capacitori “dimezzati” per colonna (2x0.5C/BL).

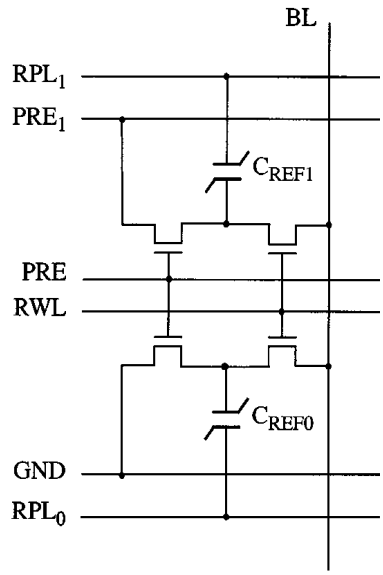


Figura 3.13. [1, 17] Schema di “Lowrey et al.” per la generazione del valore di tensione di riferimento V_{REF}.

Un secondo schema di generazione della tensione di riferimento V_{REF} è quello proposto da Lowrey et al. [17] ed appunto mostrato nella Fig. 3.16.
In primo luogo, è possibile notare come anche in questo caso la cella generatrice sia formata da due capacitori ferroelettrici di riferimento, indicati con C_{REF0} e C_{REF1}, ciascuno dei quali dotato di un proprio transistore di accesso.
Le caratteristiche principali di questi due condensatori sono: (1) entrambi hanno dimensioni capacitive pari esattamente alla metà di quelle che caratterizzano i comuni capacitori ferroelettrici C_{FE} presenti nel dispositivo; (2) in C_{REF0} e C_{REF1} sono rispettivamente memorizzati i valori logici “0” ed “1” (a differenza quindi di quanto accade invece per il primo schema di generazione della tensione di riferimento).
Questa seconda caratteristica, fa sì che nei due capacitori di riferimento siano presenti valori di tensioni circa pari a $V_0/2$ e $V_1/2$, sicché una loro contemporanea attivazione (che avviene ovviamente portando a livello alto sia il segnale di RWL, sia quelli di RPL_0 e RPL_1), fa sì che sulla Bitline sia poi presente un valore totale di tensione proprio pari a $(V_1+V_0)/2$.
In realtà, il valore di tensione di riferimento V_{REF} presente sulla Bitline è pari a:

$$V_{\mathrm{REF}} = V_{\mathrm{DD}} \left(\frac{C_0/2 + C_1/2}{C_0/2 + C_1/2 + C_{\mathrm{BL}}} \right)$$

dato appunto che i due capacitori di riferimento C_{REF0} (con capacità $C_0/2$) e C_{REF1} (con capacità $C_1/2$) quando sono connessi alla Bitline, costituiscono tra loro un parallelo, che a sua volta risulta essere disposto in serie con la capacità parassita C_{BL} della BL.
Questo valore di riferimento è in effetti più grande del valore ideale $(V_1+V_0)/2$ di V_{REF}. Ciononostante, esso è utilizzato ugualmente come valore

con cui confrontare le tensioni memorizzate nei capacitori ferroelettrici delle comuni celle di FeRAM anche se, in fase di progettazione, bisogna tener conto del fatto che il valore di V_{REF} di cui sopra è molto più prossimo al valore V_1, di quanto non lo sia a quello di V_0, cosicché il margine di immunità ai disturbi per il valore logico "1" è molto più ridotto, rispetto a quello relativo al valore logico "0".

L'altro inconveniente di questo schema è rappresentato dal fatto che il circuito di generazione di V_{REF} sia difatti attivato ogniqualvolta una cella che appartenga alla medesima colonna alla quale appartiene il circuito di riferimento in considerazione, debba appunto essere letta. Ciò implica che i condensatori ferroelettrici che costituiscono le celle di generazione di V_{REF}, siano soggetti ad un più inteso fenomeno di fatigue.

Inoltre, bisogna fare in modo che il processo di ripristino dei valori logici "1" e "0" per i condensatori C_{REF1} e C_{REF0} non determinino l'insorgenza di fenomeni di imprint.

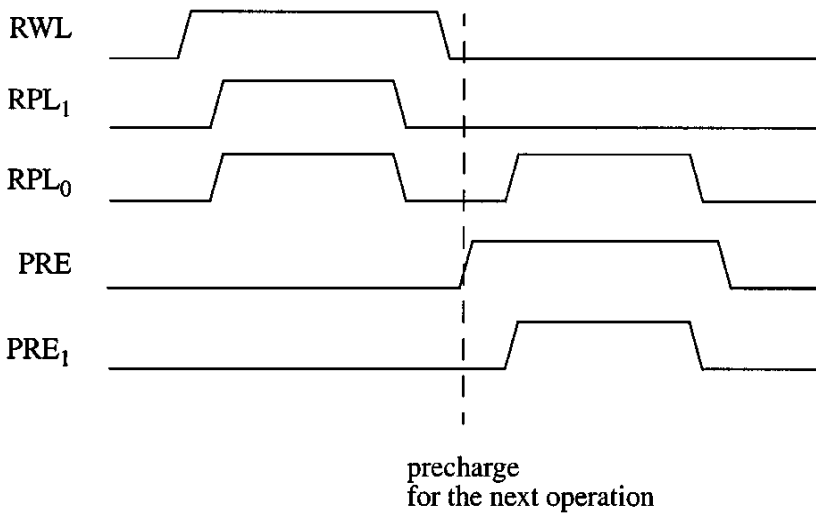


Figura 3.14. [1, 17] Diagrammi temporali dei segnali presenti sulle linee di connessione durante l'acceso in lettura ad una cella di riferimento (da notare l'andamento dei segnali PRE e PRE_1, utili appunto per il ripristino dei valori di tensione per i condensatori C_{REF0} e C_{REF1}).

3.6 – Architetture delle Memorie Ferroelettriche.

Realizzando un confronto tra le memorie DRAM e quelle FeRAM emergono numerose similarità nell'architettura circuitale interna dei dispositivi che appartengono a queste due categorie di memorie [18].

3.6.1 – Architettura con Wordline parallela alla Plateline (WL//PL).

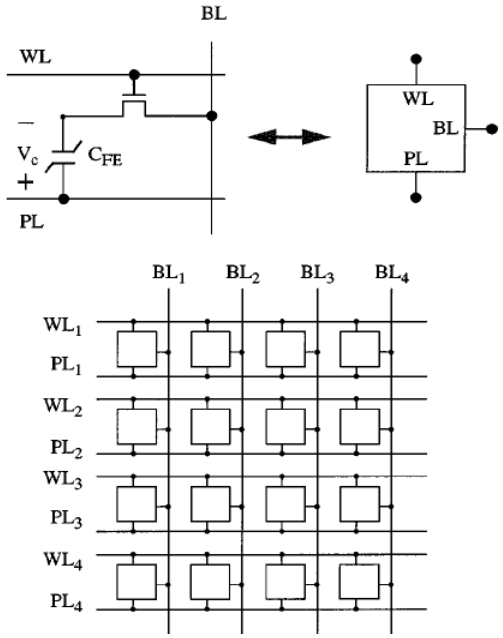


Figura 3.15. [1] Schema a blocchi di una FeRAM con architettura "WL//PL".

Come mostrato nella figura di sopra, un'architettura "WL//PL" ha come propria caratteristica principale quella di possedere le linee di PL e di WL disposte parallelamente tra loro.

Una prima deduzione riguardo allo schema circuitale di cui sopra può essere fatta notando che, l'attivazione di una coppia di linee (WL;PL),

implica un accesso simultaneo a tutte le celle che a quella riga appartengono.

Questo aspetto è soventemente ravvisabile in un qualsiasi dispositivo RAM, nonché il fatto che non si possa accedere singolarmente ad un'unica cella di memoria è di per se funzionale ad una coerente e consistente lettura della memoria, dato che nelle celle adiacenti di una medesima riga sono generalmente memorizzati ordinatamente i bit adiacenti che costituiscono i byte salvati nel dispositivo.

La condivisione della linea PL tra due righe consecutive [15] costituisce una prima e semplice modalità per ridurre l'area occupata dal dispositivo.

E' ovvio che l'adozione di questa soluzione comporta una serie di inconvenienti, primo fra tutti il fatto che le celle selezionate solo a metà (cioè aventi la linea di PL attivata, ma non quella di WL), potrebbero essere disturbate nei valori di tensione presenti ai capi dei loro capacitori ferroelettrici.

In particolare, questa anomalia si verifica perché il nodo di memorizzazione della cella presenta anch'esso una propria capacità parassita, che forma poi un “partitore di capacità” con lo stesso capacitore ferroelettrico di cella: ne consegue pertanto che il valore di tensione ai capi del singolo C_{FE} in esame può risultare alterato a causa della ripartizione di carica tra lo stesso capacitore C_{FE} e quello parassita del nodo, realizzata al momento dell'attivazione della linea di PL.

In realtà, questo fenomeno non rappresenta un disturbo nel caso in cui il valore logico precedentemente memorizzato nella cella sia uno “0”.

Al contrario, se nella cella è presente un valore logico alto, allora le tensioni che si sviluppano ai capi del C_{FE} possono invertire lo stato di bit precedentemente salvato (ai capi del C_{FE}, dopo il processo di ripartizione

di carica attivata dalla segnale sulla linea di PL, sarà presente una tensione simile a quella che rappresenta un valore logico basso di bit).
In realtà, questo fenomeno è trascurabile (è raro che si sviluppino tensioni ai capi del C_{FE} che producono la transizione [1 → 0]), anche se è possibile che il verificarsi di una successione consecutiva di questi disturbi possa determinare alla fine una transizione [1 → 0] non voluta.

3.6.2 – Architettura con Bitline parallela alla Plateline (BL//PL).

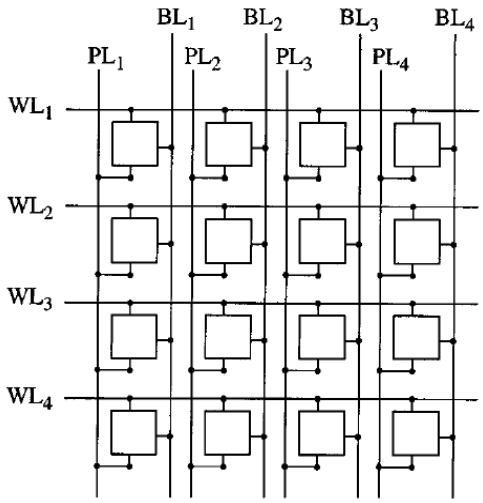


Figura 3.16. [1] Schema a blocchi di una FeRAM con architettura "BL//PL".

Nella figura precedente è mostrato lo schema a blocchi di una FeRAM con architettura "BL//PL", che per l'appunto evidenzia una disposizione parallela tra le Bitline e le Plateline delle celle di memoria [19].
A differenza di quanto accade per lo schema implementativo "WL//PL", per quest'architettura la selezione delle celle avviene singolarmente (vedi attivazione della PL y-esima, e contemporaneamente della WL x-esima, che comporta la lettura della cella che si trova in corrispondenza in posizione (x;y) nella matrice).

Inoltre, questo schema circuitale prevede anche la possibilità di eseguire la lettura di tutte le celle disposte su una riga (vedi attivazione di una singola WL e simultaneamente di tutte le PL delle colonne incluse nell'architettura a matrice della memoria).
La lettura di una singola una cella di memoria consente di ridurre notevolmente il consumo di potenza richiesto (un solo s.a. sarà attivato contestualmente). Tale consumo ovviamente aumenta nelle situazioni che comportano la lettura di un'intera riga, data appunto la maggiore potenza richiesta nella carica e scarica delle PL associate alle colonne della matrice.
In ultimo, è importante sottolineare come anche per questa architettura, le "celle selezionate a metà" (cioè quelle che apparterranno alla stessa colonna, ma non alla stessa riga di quella effettivamente da leggere) soffrono dello stesso fenomeno di disturbo dello stato di bit [19] riscontrato nelle architetture con "WL//PL" a "PL condivisa tra righe adiacenti".

3.6.3 – Architettura a Plateline segmentata (Segmented PL).

La struttura tipica di una FeRAM con architettura "Segmented PL" è mostrata nella Fig. 3.17. Questo schema implementativo è sicuramente migliore rispetto a quelli "WL//PL" e "BL//PL" dato che, a differenza di quest'ultimi, esso non comporta elevati consumi di potenza nell'accesso in lettura ad una riga di celle.

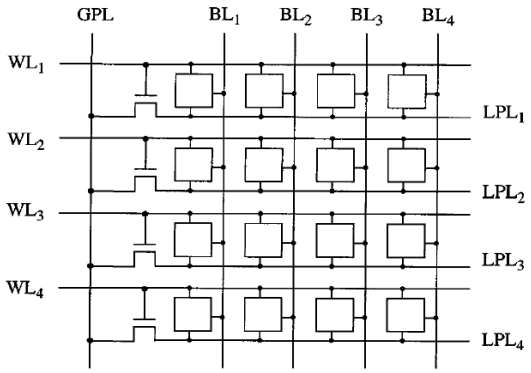


Figura 3.17. [1, 20] Schema a blocchi di una FeRAM con architettura "Segmented PL".

L'idea alla base di quest'architettura sta proprio nel fatto che il blocco di memoria dispone di un insieme di Plateline Locali (Local PL o LPL), ognuna delle quali adibita al controllo di un numero ridotto di celle di riga, garantendo così che le operazione di lettura avvengano in modo molto più rapido di quanto accade invece per il caso di un'architettura WL//PL.

Inoltre, non bisogna dimenticare che ciascuna LPL è a sua volta comandata da una Plateline generale (General PL o GPL) [16, 20], alla quale appunto risulta connessa mediante un transistore il cui gate è controllato dalla WL a cui la stessa LPL è associata.

Detto ciò, l'accesso in lettura è realizzato attivando la WL della riga di celle che si intende leggere, quindi attivando la linea GPL.

Un'architettura di questo tipo non soffre neanche del fenomeno di disturbo degli stati di bit tipico dello schema "WL//PL-condivisa" e BL//PL, visto che le LPL saranno effettivamente attivate se e solo se il transistor che le collega alla GPL risulta essere acceso.

La scelta del numero di LPL e di GPL da inserire nel blocco di memoria influenza la velocità ed il consumo di potenza del dispositivo di memoria FeRAM che adotta effettivamente questo schema.

3.6.4 – Architettura con Wordline e Plateline incorporate (Merged-Line ML).

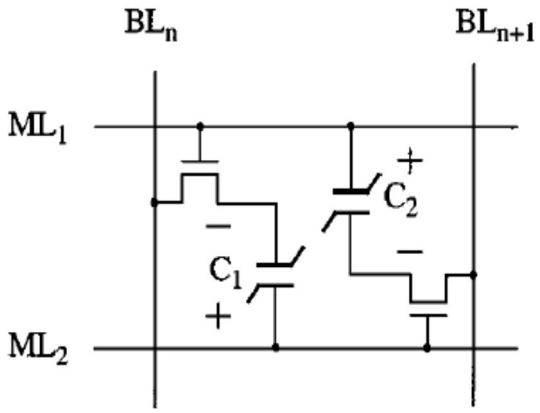


Figura 3.18. [1, 21] Coppia di celle di memoria per una FeRAM con architettura ML.

L'architettura proposta da Kang et al. [21] prevede che ciascuna coppia [Wordline;Plateline associata] sia sostituita da un'unica linea detta Merged-Line (ML). Nella Fig. 3.18 è rappresentata una coppia di celle, nella quale appunto sono ben visibili le rispettive BL (vedi BL_n e BL_{n+1}), e le corrispondenti Merged-Line M_1 e M_2 (la prima di queste sostituisce la coppia di linee [WL_1;PL_1]; mentre la seconda rimpiazza la coppia [ML_2;PL_2]).

Siccome ciascuna ML gioca il doppio ruolo di WL e PL, allora le operazioni di scrittura e di lettura dei valori logici di cella sono ovviamente più complesse [21].

Confrontando l'architettura WL//PL con quella in esame, è possibile in primo luogo rilevare come lo schema Merged-Line offra una maggiore velocità d'accesso in fase di lettura, dato appunto che la capacità associata ad una generica Plateline è difatti dimezzata ed allocata equamente alle due ML corrispondenti.

Ad esempio, per il caso mostrato in Fig. 3.18, la plateline $PL_{i\text{-}esima}$ che afferisce i capacitori ferroelettrici C_1 e C_2, appunto acceduti mediante la linea $WL_{i\text{-}esima}$ in un ipotetico schema WL//PL, confluisce nelle linee ML_1 e ML_2, assieme ovviamente alla WL_i, cosicché la capacità iniziale associata alla stessa PL_i risulta dimezzata tra le due linee di tipo merged.

Un secondo vantaggio è rappresentato dalla maggiore densità d'integrazione offerta dallo schema ML, vista appunto la notevole riduzione del numero di linee di accesso che caratterizza questo approccio (è presente una ML in luogo di una coppia [WL;PL]).

Per quanto riguarda invece gli svantaggi, è importante rilevare come il processo di fabbricazione del blocco di memoria diventa altresì ben più complesso rispetto a quelli adoperati per l'implementazione degli altri schemi architetturali descritti in precedenza [21].

3.6.5 – Architettura Non-Driven Plateline (NDP).

L'approccio ideato da Koike et al. [22] prevede che sulla Plateline di ogni cella di memoria sia presente un valore di tensione costante durante le fasi di lettura e scrittura.

Si dimostra che questa peculiarità fa sì che l'architettura NDP offra tempi di accesso in lettura/scrittura ridotti rispetto a quelli che comunemente caratterizzano gli altri schemi implementativi precedentemente analizzati [22].

Uno dei principali svantaggi di questo approccio, riguarda la necessità di ripristino dei valori logici letti, essendo appunto le operazioni di lettura "distruttive" (ciò aumenta quindi il tempo necessario per il completamento di un'operazione di lettura).

3.6.6 – Memorie Ferrolelettriche "Dual-Mode".

Uno dei problemi che maggiormente affliggono i dispositivi FeRAM, è l'occorrenza del degrado per "fatigue" dei capacitori ferroelettrici.
Una prima semplice soluzione utilizzabile per risolvere questo problema, è quella di impiegare i C_{FE} in "modalità non-switching" durante le normali operazioni di lettura/scrittura del dispositivo, realizzando appunto le operazioni di cambiamento della polarizzazione elettrica degli stessi capacitori, solamente durante le fasi di accensione/spegnimento del componente di memoria.

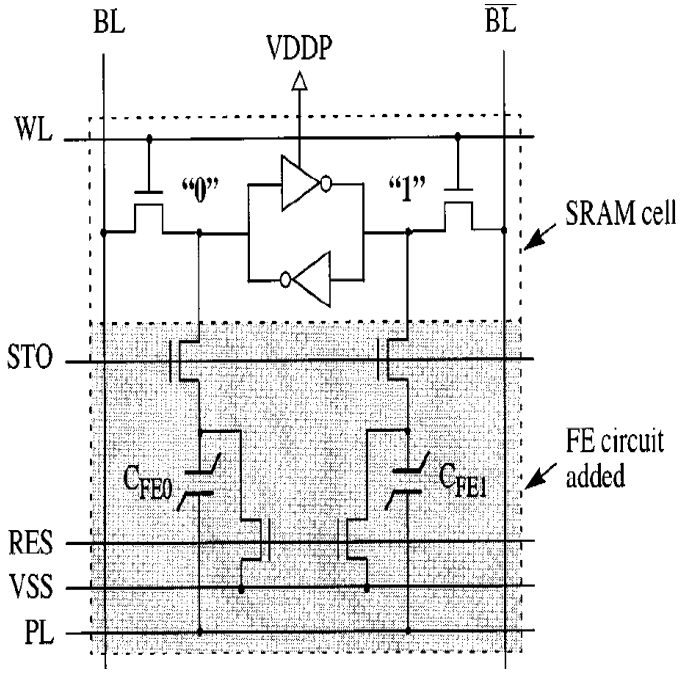


Figura 3.18. [1] Schema circuitale di una memoria Shadow SRAM.

Il modello di FeRAM noto come "Shadow SRAM" [23] è un esempio di memoria ad architettura dual-mode. Come si vede in Fig. 3.18, ogni cella di memoria è costituita da due parti: la prima, disposta nella parte

superiore, corrisponde esattamente ad una cella di SRAM; la seconda parte invece, collocata nella parte inferiore dello schema in esame, costituisce una cella di FeRAM di una particolare architettura circuitale, nota appunto come “fully-differential ferroelectric memory”.

Il funzionamento del dispositivo in esame è il seguente: le normali operazioni di lettura/scrittura dei dati avvengono utilizzando le celle di SRAM (metà superiore), sicché la restante parte con capacitori ferroelettrici è isolata e non utilizzata (la tensione presente sul segnale STO è infatti bassa durante lo svolgimento delle operazioni di cui sopra).

La metà inferiore della cella a condensatori ferroelettrici è attivata non appena inizia lo spegnimento del dispositivo, e cioè portando la linea di STO ad un livello di tensione alto, così da togliere dall'isolamento la parte fully-differential della cella di memoria.

In questa fase dunque, vengono salvati i valori logici presenti sui nodi di memorizzazione della SRAM, visto appunto che i gli stessi nodi della cella superiore fungono da bitline per i C_{FE0} ed il C_{FE1} presi in esame, mentre la linea di STO agisce come se fosse la wordline della medesima cella fully-differential.

Il salvataggio avviene generando un impulso di tensione sulla linea PL, così da poter scrivere i valori logici dei nodi di memorizzazione della SRAM all'interno dei singoli C_{FE}. Infine, lo shutdown della FeRAM avviene non appena l'impulso di tensione presente sulla linea di PL termina.

Per quanto riguarda invece la procedura di ripristino dei valori logici realizzata durante l'accensione del dispositivo di memoria, essa ha inizio portando ad un valore di tensione alto la linea di STO, quindi generando anche in questo caso un impulso di ampiezza V_{DD} sulla linea PL, cosicché

i valori logici salvati precedentemente nei capacitori ferroelettrici possano essere trasferiti sui nodi di memorizzazione della cella di SRAM.

In genere, i C_{FE} che possiedono un “1” logico producono un più alto voltaggio sul corrispondete nodo di SRAM rispetto appunto ai valori di tensione definiti dai capacitori ferroelettrici che possiedono uno “0” logico. Ciò permette alla cella superiore di raggiungere immediatamente una condizione per cui le tensioni di nodo sono a regime e rappresentano coerentemente i valori precedentemente salvati nei C_{FE}. A questo punto, la cella fully-differential viene nuovamente isolata (la tensione presente sulla linea di STO è infatti portata a 0V).

I diagrammi temporali delle procedure eseguite in caso di spegnimento e accensione del dispositivo di FeRAM sono mostrati nella Fig. 3.19.

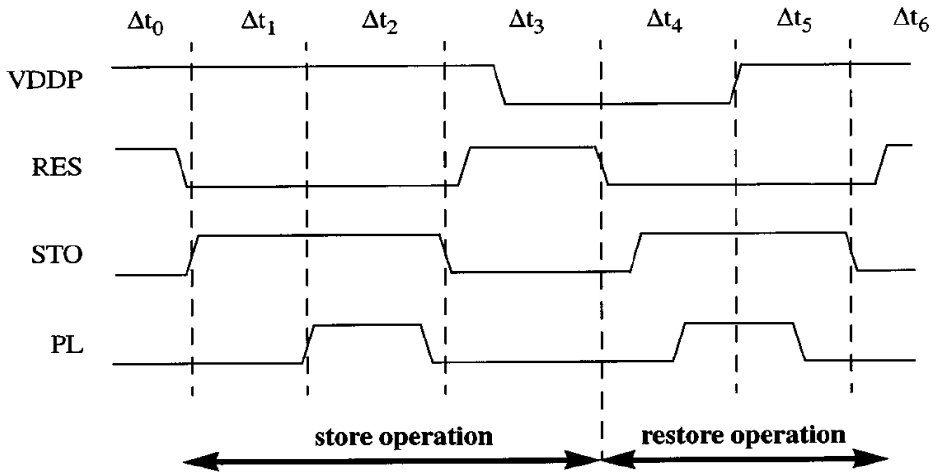


Figura 3.19. [1, 23] Diagrammi temporali delle tensioni presenti sulle linee VDDP (tensione di alimentazione), STO, PL e RES.

In ultimo, si noti come in corrispondenza degli intervalli temporali Δt_3 e Δt_6 venga attivata la linea di RES, utilizzata appunto per portare a 0V il nodo di connessione tra nMOS di accesso alla cella fully-differential, ed il C_{FE} corrispondente (in questo modo si è sicuri infatti che la tensione ai capi di

ciascun capacitore ferroelettrico è 0V, dato appunto che anche la linea di PL ha tensione a livello basso).

3.6.7 – Architettura matriciale a celle ferroelettriche amplificanti (Ferroelectric Gain Cells).

I dispositivi FeRAM con architettura matriciale a celle ferroelettriche amplificanti mostrano un'elevata densità di integrazione, dato che non necessitano ad esempio dell'utilizzo delle linee di plateline, nonché permettono l'esecuzione di operazioni di lettura “non distruttive” [24] (caratteristica non usuale per la maggior parte delle architetture FeRAM).

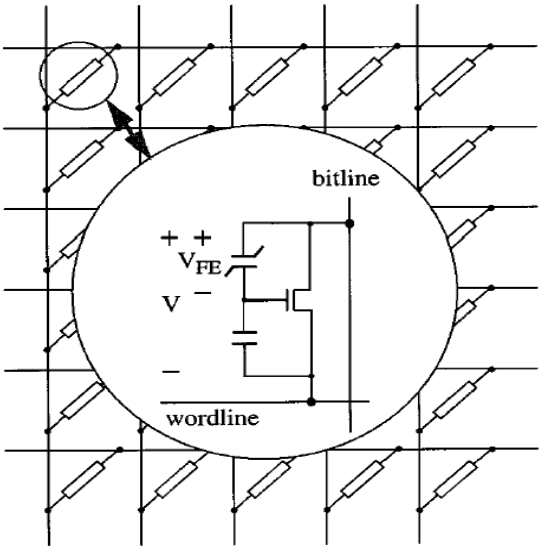


Figura 3.19. [1, 24] Architettura a matrice del dispositivo FeRAM in esame, e schema circuitale di una “Ferroelectric Gain Cell”.

Una cella ferroelettrica amplificante è costituita per l'appunto da: un condensatore ferroelettrico, un capacitore lineare (la cui occupazione di spazio è veramente marginale), ed infine un transistore. I due capacitori sono disposti in serie in modo da formare un “partitore di capacità”, la cui

tensione di uscita è per l'appunto amplificata dal transistore incluso nello schema circuitale di cella.

In particolare, il funzionamento di una FeRAM avente questa struttura circuitale prevede: (1) per la modalità stand-by, l'attivazione di tutte le bitline e di tutte le wordline (saranno appunto portate tutte a V_{DD}), in modo tale che le celle presentino ai loro capi una tensione pari a 0V; (2) per l'esecuzione invece di un'operazione di scrittura, sono generati opportuni impulsi sulle linee di BL e di WL collegate alle celle da scrivere, di modo che ai capi della "serie di capacitori di cella" risulti applicata una tensione pari a $+V_{DD}$ (WL = 0V; BL=V_{DD}) o a $-V_{DD}$ (WL = $+4V_{DD}/3$; BL = $V_{DD}/3$), a seconda appunto che si voglia salvare un valore logico alto o basso [24].

Per quanto riguarda invece le operazioni di lettura, esse cominciano precaricando la linea di BL a V_{DD}, mentre quella di WL ad una tensione di poco inferiore allo stesso valore V_{DD}. A questo punto, a seconda che nella cella sia salvato uno "0" od un "1", l'intensità di corrente che porterà a 0V la BL della cella da leggere (scaricata appunto attraverso il transistor di amplificazione verso la WL corrispondente), sarà rispettivamente più o meno elevata (cioè: $I_{READ\_"0"} >> I_{READ\_"1"}$, vedi anche caratteristica corrente-tensione mostrata in Fig. 3.20).

Pertanto, l'operazione di pulling-down di una BL nel caso in cui la cella contenga al proprio interno un "1" sarà molto più lenta di quella che si ha nel caso in cui il valore logico salvato sia "0". Ovviamente, anche in questo caso la lettura prevede in ultimo l'attivazione di un s.a..

Come si diceva all'inizio, quindi, le operazioni di lettura non richiedono l'esecuzione di una successiva procedura di ripristino dei valori logici precedentemente letti.

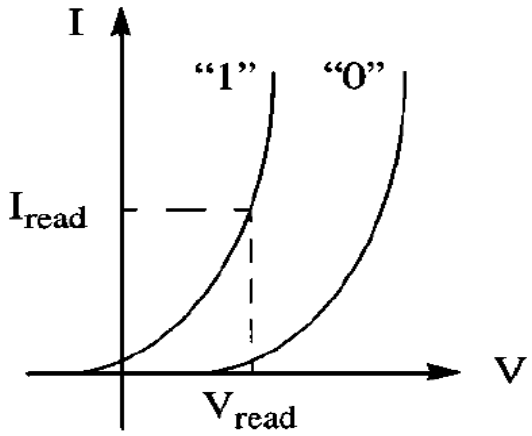


Figura 3.20. [1] Caratteristica Corrente-Tensione di una cella ferroelettrica amplificante.

3.6.8 – Architettura Chain FeRAM.

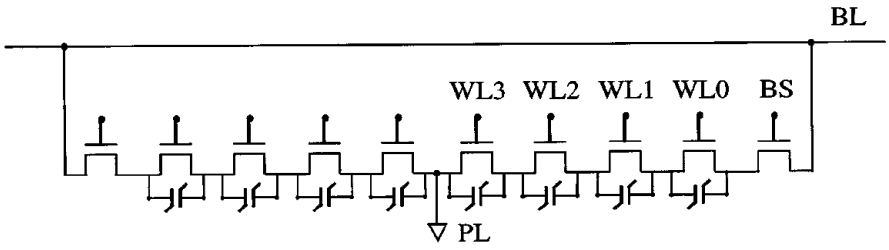


Figura 3.21. [1, 25] Schema circuitale per un dispositivo di memoria Chain FeRAM.

La caratteristica principale di un'architettura "Chain FeRAM" [25, 26] è rappresentata dalla presenza di raggruppamenti di celle di memoria all'interno dei quali le stesse sono disposte in serie tra loro: ciascuna serie ha inoltre uno proprio terminale connesso alla linea di plateline, mentre l'altro collegato alla linea di bitline (vedi Fig. 3.21).

La diminuzione del numero di contatti (verso appunto le BL e le PL) introdotta da questa architettura determina una riduzione dello spazio

occupato dal dispositivo di memoria ed un contestuale miglioramento dei tempi di accesso.

Anche questo schema circuitale contempla la modalità di funzionamento stand-by, che comporta appunto l'accensione di tutti i transistor che controllano i C_{FE}, e cioè: attivando tutte le wordline, si fa in modo che ai capi di ogni condensatore ferroelettrico vi sia una caduta di tensione nulla, visto appunto che ogni nMOS acceso cortocircuita su se stesso ciascun capacitore.

Nel caso invece si voglia realizzare la lettura di una singola cella, si procede allora disattivando solo la WL del C_{FE} che si intende leggere, ed altresì si lasciano attive le restanti wordline. In questo modo, infatti, la tensione presente sulla linea di PL raggiunge l'unico condensatore ferroelettrico non cortocircuitato attraverso la catena di transitori, ed il valore logico posseduto dallo stesso C_{FE} raggiunge la linea di bitline corrispondente per mezzo della restante parte di catena. Ovviamente, l'operazione preliminare a tutto ciò è quella di accendere il transistor comandato dal segnale di BS (che quindi sarà ad un livello alto di tensione).

Un aspetto vantaggioso di quest'approccio è rappresentato dal fatto che la capacità parassita sentita sulla BL avrà valori molto più contenuti rispetti a quelli tipici degli approcci 1T-1C, dato appunto che in quest'architettura risulta connessa direttamente alla stessa BL solamente una cella.

Inoltre, questa architettura consente di allocare un numero maggiore di celle per Bitline, ovvero permette la realizzazione di blocchi di serie con un numero di celle N piuttosto elevato.

In realtà, un eccessivo incremento di quest'ultimo parametro, determina un aumento della resistenza e della capacità parassite associate a ciascuna

serie [25], andando così poi ad incidere negativamente sui ritardi riscontrabili durante le fasi di lettura e scrittura dei valori logici.

E' necessario quindi scegliere un valore per il parametro “numero di celle per serie” che permetta di ottenere una buona riduzione delle dimensioni del chip e che contemporaneamente non determini un aumento indiscriminato della capacità e resistenza di blocco che infici poi eccessivamente il ciclo di readout del dispositivo.

Un ulteriore svantaggio associato a quest'approccio riguarda il fatto che i valori logici salvati nelle celle non selezionate durante le operazioni di lettura/scrittura potrebbero essere alterate a causa della resistenza parassita del transistor della wordline corrispondente (questo problema si risolve scegliendo un valore per il parametro N che sia uguale o superiore a 16 celle per blocco [25, 26], e realizzando delle procedure di “rinforzo” del valore logico letto o scritto).

Capitolo 4 – Memorie a cambiamento di fase.

Precisazioni.

Il seguente capitolo è stato scritto riassumendo e traducendo in parte l'articolo [1], dal quale inoltre sono state tratte anche le figure ed i grafici che ivi compaiono.

Le memorie a cambiamento di fase (PCM, Phase Change Memory) offrono meccanismi di memorizzazione di tipo non-volatile, che si dimostrano essere compatibili con una progressiva riduzione delle dimensione dei processi produttivi delle celle.
In particolare, tenendo presente che il funzionamento delle PCM si basa sul controllo di opportune correnti e sulla realizzazione di effetti termici che consentono di modificare la "fase" degli elementi di memoria di ciascuna cella, e che inoltre lo scaling della tecnologia PCM e delle relative aree di contatto controllate termicamente, determina una corrispondente riduzione delle intensità di corrente richieste in fase di programmazione delle celle; allora è possibile dimostrare che processi processi produttivi di 20nm sono effettivamente utilizzabili per l'implementazione di dispositivi PCM (si stima

addirittura che sia possibile raggiungere la soglia dei 9nm [5, 23]).

Tuttavia, sebbene le PCM offrano elevate capacità di memorizzazione dati e garantiscano un'alta densità di integrazione, dall'altra parte questi componenti evidenziano aspetti negativi che contrariamente non affliggono i restanti dispositivi di memoria che attualmente sono presenti sul mercato.

E cioè, in primo luogo le PCM mostrano elevati tempi di latenza nell'accesso in lettura/scrittura rispetto a quelli che caratterizzano le memorie DRAM, le quali pertanto risultano essere nettamente più veloci rispetto alle stesse PCM.

In seconda battuta, questi componenti di memoria presentano alti consumi energetici durante l'esecuzione delle proprie operazioni di scrittura, ed in aggiunta mostrano rapidi degradi delle proprie celle di memoria (nota: essi sono determinati dalle sollecitazioni per effetto termico subite dagli elementi di memoria, prodotte per l'appunto dalle stesse operazioni di programmazione delle celle).

Inoltre, proprio l'ultimo aspetto di cui sopra incide negativamente sul numero di operazioni di programmazione di cella che è possibile eseguire su un generico dispositivo PCM (vedi parametro di “Endurance”).

Gli svantaggi elencanti di sopra, rappresentano le ragioni per cui questa tecnologia realizzativa occupa in realtà solo fette marginali del mercato delle memorie non-volatili, nonché nel caso si volesse utilizzare le stesse come valide alternative alle memorie DRAM, allora sarebbe necessario realizzare un ripensamento delle PCM in esame, così da renderle utilizzabili come memorie principali all'interno di architetture general-purpose.

	Horri [11]	Ahn [2]	Bedeschi [6]	Oh [20]	Pellizer [21]	Chen [8]	Kang [12]	Bedeschi [7]	Lee [15]	Parameters [this work]
Year	2003	2004	2004	2005	2006	2006	2006	2008	2008	**
Process (nm, F)	**	120	180	120	90	**	100	90	90	90
Array Size (Mb)	**	64	8	64	**	**	256	256	512	**
Material	GST, N-d	GST, N-d	GST	GST	GST	GS, N-d	GST	GST	GST	GST,N-d
Cell Size (μm^2)	**	0.290	0.290	**	.097	60 sq-nm	0.166	0.097	0.047	0.065-0.097
Cell Size (F^2)	**	20.1	9.0	**	12.0	**	16.6	12.0	5.8	9.0-12.0
Access Device	**	**	BJT	FET	BJT	**	FET	BJT	diode	BJT
Read T (ns)	**	70	48	68	**	**	62	**	55	48
Read I (uA)	**	**	40	**	**	**	**	**	**	40
Read V (V)	**	3.0	1.0	1.8	1.6	**	1.8	**	1.8	1.0
Read P (uW)	**	**	40	**	**	**	**	**	**	40
Read E (pJ)	**	**	2.0	**	**	**	**	**	**	2.0
Set T (ns)	100	150	150	180	**	80	300	**	400	150
Set I (uA)	200	**	300	200	**	55	**	**	**	150
Set V (V)	**	**	2.0	**	**	1.25	**	**	**	1.2
Set P (uW)	**	**	300	**	**	34.4	**	**	**	90
Set E (pJ)	**	**	45	**	**	2.8	**	**	**	13.5
Reset T (ns)	50	10	40	10	**	60	50	**	50	40
Reset I (uA)	600	600	600	600	400	90	600	300	600	300
Reset V (V)	**	**	2.7	**	1.8	1.6	**	1.6	**	1.6
Reset P (uW)	**	**	1620	**	**	80.4	**	**	**	480
Reset E (pJ)	**	**	64.8	**	**	4.8	**	**	**	19.2
Write Endurance	1E+07	1E+09	1E+06	**	1E+08	1E+04	**	1E+05 (MLC)	1E+05	1E+08

Tabella 4.1. [1] Confronto tra parametri implementativi e di funzionamento dei prototipi di memoria PCM realizzati dal 2003 al 2008. Nell'ultima colonna sono mostrati i valori medi dei parametri esaminati, ed appunto determinati durante l'analisi condotta dagli autori dell'articolo [1].

4.1 – Celle di memoria di un dispositivo Phase-Change Memory.

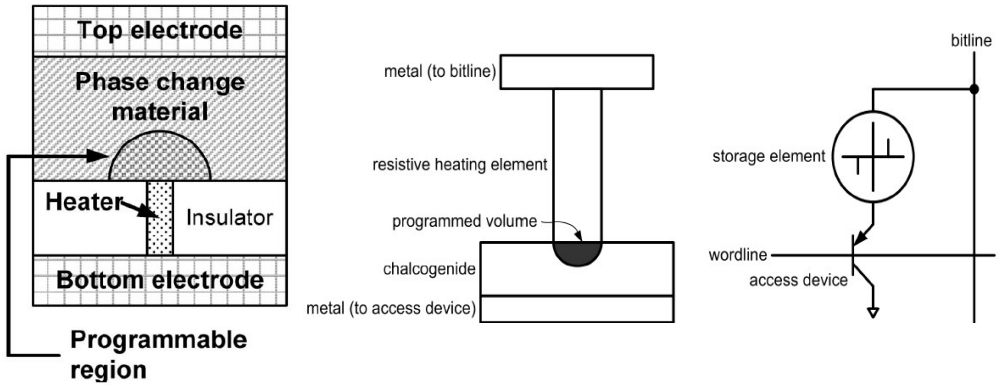


Figura 4.2. Sezione verticale di un generico elemento di memoria di una cella appartenente ad un dispositivo PCM (immagine a sinistra [3]) e sezione con relative indicazioni delle connessioni dei due elettrodi dello stesso componente verso le linee di Bitline e Wordline (immagine centrale [1]). Nella figura a destra [1] invece, è rappresentata un'intera cella di una Phase-Change Memory, avente come

proprio elemento di accesso un BJT.

Come mostrato nella figura di sopra, ciascun elemento di memoria è costituito da due elettrodi metallici (vedi Top electrode e Bottom electrode), separati da un sottile strato che: per metà risulta essere realizzato in materiale calcogenico (in questa sottile pellicola è allocata anche la regione programmabile); e per l'altra metà invece è realizzato in materiale isolante (all'interno dello stesso è ricavato l'elemento resistivo necessario per compiere la scrittura della singola cella).

Il composto calcogenico o calcogenuro (nota: un composto calcogenico è un materiale che include nella propria struttura cristallina atomi di Ossigeno e/o Zolfo e/o Selenio e/o Tellurio e/o Polonio [4]) più frequentemente adoperato per l'implementazione di celle dei chip di memoria PCM è il $Ge_2Sb_2Te_5$ (GST).

In realtà esistono e sono utilizzati anche altri calcogenuri, dato appunto che spesso questi altri tipi di materiali calcogenici offrono più alti valori di resistività e migliori caratteristiche elettriche rispetto a quelle che tipicamente caratterizzano il semplice GST.

A tal riguardo, un esempio è quello del calcogenuro GST drogato con atomi di Azoto, che per l'appunto presenta più alti valori di resistività rispetto a quelli del semplice GST, nonché permette di utilizzare valori più piccoli d'intensità di corrente durante la fase programmazione dell'elemento di memoria.

La programmazione di un elemento di memoria avviene facendo scorrere della corrente nella giunzione [strato calcogenico-strato con resistenza], cosicché la testina (vedi figure precedenti) in corrispondenza dell'area di contatto tra le due precedenti pellicole, raggiunge temperature pari a circa 650°C, realizzando così un cambiamento di fase (si passa da struttura cristallina ad amorfa o viceversa in base all'ampiezza ed alla durata

dell'impulso di corrente).

La caratteristica corrente-tensione delle testine di memoria in materiale calcogenico non dipende dallo stato iniziale in cui le stesse si trovano, pertanto la complessità ed i tempi di latenza delle operazioni di scrittura delle celle PCM divengono piuttosto ridotti [14].

Come per tutti i dispositivi di memoria, l'accesso ai componenti di memorizzazione può avvenire utilizzando alternativamente: un BJT, un MOSFET, o ancora un diodo.

In realtà, il primo di questi elementi circuitali è quello maggiormente utilizzato poiché permette di bypassare i punti deboli tipici dei MOSFET e dei diodi per l'appunto, consentendo così un accesso veloce alla singola cella ed offrendo al contempo la possibilità di raggiungere un buon grado di densità d'integrazione per il dispositivo di memoria [7, 22].

4.2 – Operazioni di scrittura di una generica cella di PCM.

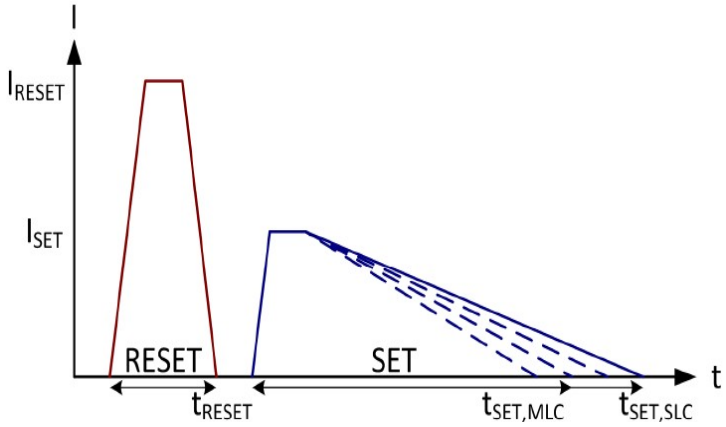


Figura 4.3. [1] Curve "Intensità di Corrente-Durata Temporale dell'Impulso di Corrente" utilizzate durante la scrittura di un valore logico basso (operazione di RESET, curva in rosso) o di un valore logico alto (operazione di SET, curva in blu). Si noti, come la durata dell'impulso di SET sia determinante per la definizione del numero di stati logici salvabili all'interno di ciascuna cella (vedi anche SLC, cioè Single Level Cell, cui corrisponde la possibilità di memorizzare 1bit/componente memoria; e MLC, cioè Multi Level Cell, al quale corrisponde la possibilità di memorizzare Nbit/componente, in genere 2bit/componente).

Le operazioni di scrittura dei valori logici “0” ed “1” avvengono adoperando impulsi di corrente di ampiezza e durata temporale differenti. In particolare, nel caso in cui si volesse eseguire un'operazione di RESET, e cioè salvare nell'elemento di memoria un valore logico basso, allora si renderebbe necessario l'utilizzo di un impulso che abbia un valore elevato di intensità di corrente, ma una durata temporale piuttosto ridotta.
In particolare, procedendo in questo modo la testina calcogenica subisce un forte sbalzo di temperatura che, vista appunto la veloce transizione da temperature elevate (causate dal flusso di corrente nella giunzione [elemento resistivo-calcogenuro]) a temperature basse (corrispondenti all'assenza di flussi di corrente), lasciano così l'elemento in calcogenuro in uno stato amorfo al quale corrisponde un elevato valore di resistenza elettrica.
Al contrario, l'esecuzione di un'operazione di SET consente di salvare nella testina di strato calcogenico un valore logico alto dato che, applicando alla giunzione per un tempo prolungato un flusso di corrente che abbia un'intensità non eccessivamente elevata, è possibile allora realizzare un graduale raffreddamento del calcogenuro, che lascia lo stesso in uno stato cristallino e non amorfo, cui corrisponde appunto un basso valore di resistenza elettrica.
Alla luce di quanto detto precedentemente, è evidente che le operazioni di SET sono quelle che maggiormente incidono sui tempi di latenza in scrittura del dispositivo PCM. In particolare, gli autori dell'articolo mostrano che con la loro procedura di estrapolazione (analizzano un processo produttivo di 180nm per determinare i parametri tipici di funzionamento di una tecnologia di fabbricazione a 90nm), l'intensità di corrente e la tensione sviluppate attraverso la giunzione ed ai capi del componente durante l'esecuzione di una tipica fase di SET, sono circa pari a 150μA ed

a 1.2V, nonché ciascuna operazione occupa un intervallo temporale di circa ben 150ns.

Se quindi in media la potenza richiesta per la scrittura di un “1-logico” è pari a 90µW, allora il dispendio energetico che ne consegue è pari a 13.5pJ (=90µW*150ns).

Per quanto riguarda invece le operazioni di RESET, anche queste mostrano un considerevole dispendio energetico. Ovvero, assumendo che il tempo richiesto, l'intensità di corrente che attraversa il componente e la tensione che si sviluppa ai capi del medesimo durante la scrittura di valore logico basso siano rispettivamente pari a 40ns [6], 300µA [6, 7] e 1.6V [6, 7], allora la potenza dissipata è di 480µW, cui corrisponde un dispendio energetico pari a 19.2pJ.

In realtà, come mostrato in tabella, esistono nuove tecnologie implementative di PCM che consentono di raggiungere tempi di latenza per un'operazione di SET inferiori ai 150ns indicati di sopra [8, 11]. Ed ancora, è possibile notare come dispositivi aventi maggiore densità di integrazione presentano in realtà tempi di scrittura considerevolmente più elevati rispetto ai 150ns di riferimento [12, 15, 20].

Il grafico delle curve “Corrente-Durata Impulso” permettono di fare un'ulteriore osservazione, e cioè le PCM rendono possibile il salvataggio di più bit logici per singolo componente di memoria.

In questi casi si parla di Multi-Level Cells [7, 19], cioè di celle che utilizzano componenti di memoria all'interno dei quali è possibile realizzare per ciascuno di essi più stati resistivi intermedi, sicché nel singolo elemento possono essere effettivamente memorizzati più bit contemporaneamente.

La realizzazione degli stati resistivi intermedi avviene notando che: mentre impulsi di SET di durata temporale elevata (vedi deboli pendenze per il fronte di discesa dell'impulso) producono transizioni parziali cui

corrispondono bassi valori di resistenza per la testina calcogenica; al contrario, fasi di SET con pendenze accentuate per il fronte di discesa dell'impulso, comportano cambiamenti di fase anch'essi parziali che possono invece coincidere con valori più o meno elevati di resistenza.
Tuttavia, la difficoltà nell'individuare e riconoscere i differenti stati resistivi presenti in un medesimo elemento di memoria, limita l'utilizzo di dispositivi Multi-Level Cell (in genere in uno stesso componente sono memorizzati al più due bit).

4.3 – Write Endurance per le memorie PCM.

Per i dispositivi Phase-Change Memory, il processo sicuramente più importante ed al contempo anche più problematico, è quello di scrittura degli elementi di memoria.
La principale fonte di problemi è rappresentata dall'immissione di corrente nella giunzione [componente resistivo-strato materiale calcogenico] che, causando effetti termici che a loro volta determinano contrazioni e/o espansioni dei materiali dell'elemento di memoria, determina così un graduale degrado delle aree di contatto tra l'elettrodo resistivo ed il calcogenuro.
Siccome pertanto il reale flusso di corrente può differire da quelli teorici ipotizzati, implicando difatti un diverso valore di resistenza elettrica associata al componente, si ha di conseguenza un'alterazione del parametro di Read Window, definito appunto come la differenza fra la massima e la minima resistenza programmabile per i componenti di memoria.
Esattamente quindi come per le memorie Flash, anche ai dispositivi PCM è possibile attribuire un valore al parametro di Write Endurance (w.e.),

indicante appunto il numero massimo di cicli di scrittura eseguibili sul singolo dispositivo, prima che esso si rompa o non possa essere più programmato in modo affidabile.

Per le PCM, l'intervallo di valori a cui appartiene questo parametro è [1E+04;1E+9] scritture, nonché il valore medio di w.e. è pari circa a 1E+07/+08 scritture (esso è evidentemente maggiore dei valori di w.e. che caratterizzano le memorie Flash, appunto in media pari a 1E+05 cicli di scrittura [13]).

A tal proposito, l'ITRS (International Technology Roadmap of Semiconductor) prevede la realizzazione di prototipi PCM con processo produttivo a 32nm e con w.e. pari a 1E+12 cicli di scrittura [5].

4.4 – Operazioni di lettura di una generica cella di PCM.

Le operazioni di lettura cominciano con una “Fase0” di precarica, durante la quale: (1) la Bitline della cella in esame è portata ad un determinato livello di tensione, indicato in genere con il nome di “Tensione di Lettura”; (2) la Wordline è mantenuta invece ad un valore basso di tensione, essendo un BJT il componente di accesso della cella.

A questo punto, a seconda che il componente di memoria sia in fase cristallina o amorfa, nonché la resistenza ad esso associata abbia un valore basso o alto, allora l'intensità di corrente che permetterà di portare a 0V la BL sarà rispettivamente più o meno elevata, e conseguentemente il transitorio di scarica della tensione di lettura sarà più veloce o più lento.

Il tempo di latenza in genere associato ad un'operazione di lettura è pari approssimativamente a 48ns [6] (in questo intervallo temporale sono inclusi la “Fase0”, l'accesso alla cella e l'esecuzione della fase di sensing della corrente/tensione presente sulla linea di BL da parte di un s.a.).

Per quanto riguarda poi la corrente e la tensione di lettura, essi assumono generalmente valori pari rispettivamente a 40μA e 1.0V, da cui si ottiene una potenza dissipata di 40μW, che nel tempo produce un dispendio energetico pari a circa 2pJ (in realtà 1.92pJ).
Infine, la scelta del BJT come dispositivo di accesso alla cella è determinante per la riduzione dei tempi di latenza (vedi il caso di utilizzo di un MOSFET o un semplice diodo, cui corrispondono tempi pari a 55 o a 70ns).

4.5 – Processo di Scaling dei dispositivi PCM.

La miniaturizzazione dei dispositivi PCM consente di utilizzare durante la fase di programmazione del medesimo intensità di corrente piuttosto piccole (la diminuzione delle aree di contatto [elettrodo resistivo-componente di memoria] e la riduzione del volume della testina di materiale calcogenico, fanno registrare difatti un aumento della resistenza elettrica del generico elemento di memorizzazione).
E cioè, la Legge di Scaling [22] individuata da Pirovano (vedi Fig. 4.4) indica che: un decremento pari a K delle dimensioni tipiche di un dispositivo PCM determina una corrispondente riduzione quadratica ($1/K^2$) dell'area di contatto di cui sopra, che a sua volta implica un aumento lineare K della resistenza di giunzione ed una diminuzione pari a (1/K) della corrente richiesta in fase di programmazione.
La “Legge di Pirovano” mostra quindi che un processo di scaling determina la riduzione delle correnti da utilizzare in fase di SET e RESET dell'elemento, ma al contempo non comporta né alcuna diminuzione delle tensioni che si sviluppano ai capi della giunzione di memoria [22], né tanto meno una riduzione dei tempi di latenza delle fasi di lettura/scrittura.

In realtà, la realizzazione di processi di scaling eccessivamente accentuati determinano in genere l'insorgenza di alcuni problemi, come ad esempio la manifestazione del fenomeno dell'accoppiamento termico tra componenti di memoria appartenenti a celle adiacenti.

E cioè, la programmazione di un elemento di memoria può influenzare lo stato di bit dei componenti che si trovano nelle vicinanze del primo. Tuttavia, lo studio condotto da Lai, evidenzia in realtà che questo problema è trascurabile [14], se si considera che la temperatura decresce esponenzialmente con l'aumentare della distanza il tra componente programmato e gli elementi che si trovano nelle sue vicinanze.

In ultimo, un altro "problema di scaling" è quello relativo al fatto che l'aumento della resistenza elettrica per area (intendendo per quest'ultima sempre quella di contatto [elettrodo resisitvo-testina materiale a cambiamento di fase]) può determinare un'eccessiva diminuzione della differenza di resistenza elettrica esistente tra fase cristallina e la fase amorfa del materiale calcogenico.

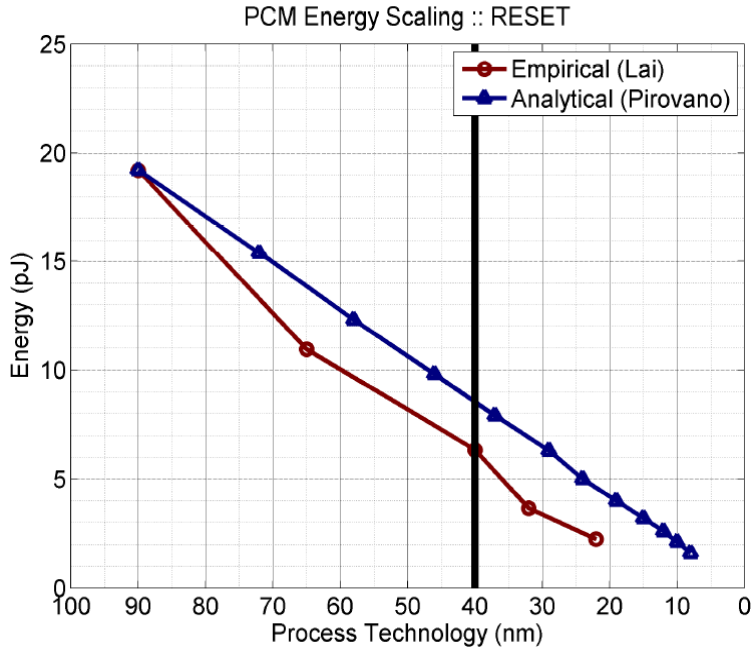


Figura 4.4. [1] Grafici che mostrano la dipendenza esistente tra la riduzione dell'energia richiesta per un'operazione RESET (asse delle ordinate) e la diminuzione delle dimensioni dei dispositivi PCM

(asse delle ascisse). In particolare, La curva in rosso, mostra il legame "teorico" esistente tra questi due parametri rilevato da Pirovano [22]. La curva in blu, invece, mostra la relazione esistente tra queste due grandezze evidenziata durante l'analisi sperimentale condotta da Lai su più prototipi PCM [14]. Si noti, come i risultati rilevati da Lai confermano la legge di scaling di Pirovano [14, 22].

4.6 – Architettura matriciale delle memorie PCM.

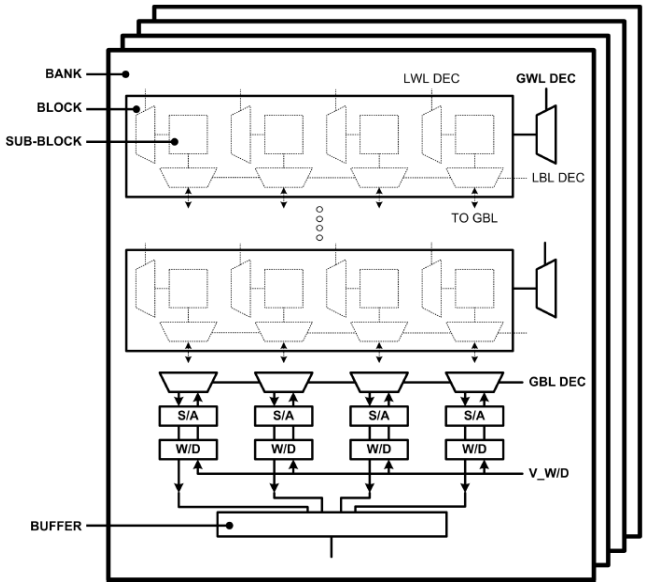


Figura 4.5. [1] Architettura di un dispositivo PCM. Nell'immagine è possibile notare la disposizione dei componenti all'interno della gerarchia Banchi → Blocchi → Sotto-Blocchi. La figura mostra inoltre come i dispositivi di sense ampliefiers e write drivers (inclusi in un generico banco) siano condivisi tra blocchi appartenenti allo stesso banco.

Come si vede nella figura precedente, anche le celle di memoria dei dispositivi PCM sono organizzate secondo un'architettura Banchi → Blocchi → Sotto-Blocchi, nonché ogni banco presenta dei propri decoder di riga e colonna (fondamentali ovviamente per l'accesso alla matrice di celle) e dei propri dispositivi di sense amplifier (s.a) e write driver (quest'ultimi sono condivisi tra tutti gli elementi di blocco presenti nel banco che si considera).

Il primo oggetto di discussione circa l'architettura di memoria, riguarda la scelta del tipo di sense amplifier da utilizzare, e cioè se è più conveniente utilizzare s.a. che si basano sul confronto di tensioni, o s.a a confronto di

corrente.

In genere quest'ultima tipologia di s.a. è quella più frequentemente utilizzata poiché i dispositivi di questa categoria realizzano la fase di sensing senza attendere che le capacità tipiche delle linee di BL si siano appunto scaricate (come invece richiesto per il caso di un confronto realizzato con un s.a. a tensione ed avente struttura ad invertitori cross-coupled), nonché risultano nettamente più veloci dei primi sopra elencati.

Per quanto riguarda le operazioni di lettura/scrittura delle righe della matrice di memoria, queste fasi avvengono sempre salvando i bit da leggere/da scrivere all'interno di uno o più buffer presenti nel banco di memoria (vedi figura di sopra).

A questo punto, confrontando le memorie DRAM con i dispositivi PCM, si nota: (1) che le operazioni di lettura per una DRAM sono distruttive, a differenza invece di quanto accade per i dispositivi PCM; (2) che le operazioni di “sensing” e “buffering” per le DRAM sono realizzate dagli stessi dispositivi di sense amplifier, al contrario invece di quanto accade per le PCM, i cui s.a. possiedono il controllo appunto di veri e propri banchi di latch utilizzati per il buffering delle righe di memoria lette.

In particolare, il fatto che nelle memorie PCM le operazioni di sensing e di buffering siano separate, implica che i s.a. siano in realtà mutiplexati, nonché nel dispositivo è prevista la presenza di decoder locali di WL e decoder locali di BL che permettono di attivare contemporaneamente righe e colonne della matrice, appartenenti però a differenti sotto-blocchi.

Quest'ultimo aspetto consente di utilizzare buffer di capacità nettamente inferiori rispetto a quelli che si dovrebbe avere se si salvassero tutti i valori logici presenti su tutte le Bitline della matrice.

4.7 – Confronto tra implementazioni standard di PCM e DRAM.

	PCM	DRAM
Delay & Timing		
tRCD (cy)	22	5
tCL (cy)	5	5
tWL (cy)	4	4
tCCD (cy)	4	4
tWTR (cy)	3	3
tWR (cy)	6	6
tRTP (cy)	3	3
tRP (cy)	60	5
tRRDact (cy)	2	3
tRRDpre (cy)	11	3
Energy		
Array read (pJ/bit)	2.47	1.17
Array write (pJ/bit)	16.82	0.39
Buffer read (pJ/bit)	0.93	0.93
Buffer write (pJ/bit)	1.02	1.02
Background power (pJ/bit)	0.08	0.08

Tabella 4.6. [1] Confronto tra parametri circuitali di una PCM e di una DRAM.

La tabella 4.6 mostra il confronto tra i parametri di funzionamento di una SDRAM DDR2 da 512 Mb prodotta da Micron (i valori sono stati individuati analizzando il datasheet del dispositivo) [16], con quelli di una PCM i cui valori degli analoghi parametri sono stati ottenuti invece prendendo come riferimento quelli evidenziati nella Tabella 4.1.

Il primo parametro temporale mostrato in tabella è quello di tRCD. Esso rappresenta il ritardo evidenziato tra la lettura della riga della matrice di memoria, e la successiva sequenza di “scrittura verso e lettura da” uno dei buffer presenti nel dispositivo. La tabella evidenzia per una memoria PCM un tRCD pari a circa 60ns (questi comprendono 48ns per realizzare la fase di lettura, e 7.5ns per la decodifica [15] della riga da leggere).

A questo punto, supponendo che la frequenza di clock utilizzato dal

dispositivo SDRAM sia pari 400MHz, allora un'operazione di lettura di una PCM occuperebbe corrispondentemente ben 22 (= 400MHz*(48+7.5)ns) cicli del clock di sopra, e cioè un numero di cicli 4.4 (= 22cycles/5cycles) volte più grande del numero di cicli richiesto da una SDRAM per eseguire in modo completo una fase di lettura.

Nelle righe successive, invece sono indicati i valori assunti dai parametri tCL, tWL, tCCD e tWTR. Questi tempi sono in realtà indipendenti dalla tecnologia implementativa delle celle del dispositivo e rappresentano solamente dei vincoli di tempo riguardanti l'esecuzione di comandi da parte dei buffer interni.

Anche i parametri tWR e tRTP non dipendono dalla tecnologia con cui sono realizzate le singole celle di memoria, ed indicano i tempi richiesti affinché un'operazione di scrittura [buffer interni → riga della la matrice di celle] sia effettivamente completata con successo (nota: l'operazione di scrittura è considerata come completata solamente quanto i valori logici presenti nelle celle risultano permanentemente stabili nel tempo).

Nella terzultima riga della tabella 4.6 (parte superiore della tabella) è mostrato il numero di cicli di clock occupati dal parametro tRP. Esso rappresenta il numero di cicli che devono trascorrere tra l'istante di tempo in cui avviene la scrittura di un insieme di celle di memoria, e quello in cui i medesimi dati sono disponibili per una nuova lettura. Nel caso di una PCM, il valore assunto da tRP è proprio pari a 60 (= 400MHz*150ns) cicli (sempre in riferimento alla frequenza di clock del dispositivo SDRAM), dato che nel calcolo di questo parametro si assume come valore del tempo di latenza di una fase di scrittura quello di un'operazione di SET (nota: si sceglie il tempo di latenza di un'operazione di SET, poiché il valore assunto da questo parametro temporale è nettamente maggiore di quello di un'operazione di RESET).

Infine, la penultima e l'ultima riga indicano i valori dei parametri temporali tRRDact (valido per gli accessi in lettura) e tRRDpre (valido per gli accessi in scrittura).
Questi parametri indicano il numero di cicli di clock che devono intercorre tra un accesso in lettura/scrittura ed uno successivo (sempre rispettivamente in lettura/scrittura), affinché il numero complessivo di accessi in un intervallo di tempo più o meno esteso non determini costi energetici superiori alla potenza fornita al dispositivo mediante la tensione di alimentazione e definita in base alla sua corrente di lavoro.

4.8 – Analisi dei consumi energetici mostrati nella Tabella 4.6.

I datasheet e le note tecniche del dispositivo SDRAM Micron preso in esame, mostrano un costo energetico pari a 1.56pJ/bit associabile ad una esecuzione in successione delle fasi di lettura e scrittura.
Da questo dato, quindi, non si comprendere quale sia l'entità del "dispendio energetico" proprio delle singole operazioni di lettura e scrittura (si assume infatti il principio che un'operazione di lettura per una DRAM è distruttiva, e dunque richiede sempre una successiva operazione di scrittura utile al ripristino dei valori logici precedentemente letti).
Tuttavia secondo l'indagine condotta nell'articolo di riferimento, l'energia richiesta per un'operazione di lettura è nettamente maggiore di quella utile per realizzare una di scrittura, nonché si stima che dei complessivi 1.56pJ/bit: 1.17pJ/bit sono attribuibili alla prima delle due operazioni di cui sopra, mentre i restanti 0.39pJ/bit sono associabili per l'appunto ad un'operazione di scrittura.
Per quanto riguarda i costi energetici di un'operazione di lettura di un dispositivo PCM invece, la Tabella 4.1 mostra un valore per questo

parametro pari a 2pJ/bit, a cui poi bisogna poi sommare il consumo dei dispositivi periferici, che per l'appunto è stimato in 0.5pJ/bit [18].

Ad una singola operazione di scrittura di una PCM è invece attribuibile un costo energetico medio pari a 16.35pJ/bit (è la media dei costi per la scrittura di uno “0-logico” e di un “1-logico”, rispettivamente pari a 13.5pJ/”0”-bit e 19.2/”1”-bit), a cui bisogna anche in questo caso aggiungere gli 0.53pJ/bit dei dispositivi periferici.

Per quanto riguarda i dispendi energetici delle fasi di lettura/scrittura dei buffer interni, i dispositivi PCM e SDRAM mostrano stessi valori per queste grandezze dato che i meccanismi e le implementazioni di questi dispositivi sono analoghi per entrambe le due tipologie di memoria considerati, nonché indipendenti dalla tecnologia realizzativa delle singole celle di memoria.

Infine, il parametro di Background-Power rappresenta il costo energetico per bit (in questi casi pari appunto 0.08pJ/bit per ogni ciclo di memoria) attribuibile ai dispositivi periferici interni delle memorie a confronto, quando in effetti le stesse memorie sono attive e pronte ad eseguire tutti i comandi [17] impartiti dal sistema che le utilizza.

4.9 – Valutazione di una memoria PCM avente architettura [1X2048B] utilizzata come Memoria Principale in sostituzione di un dispositivo standard DRAM.

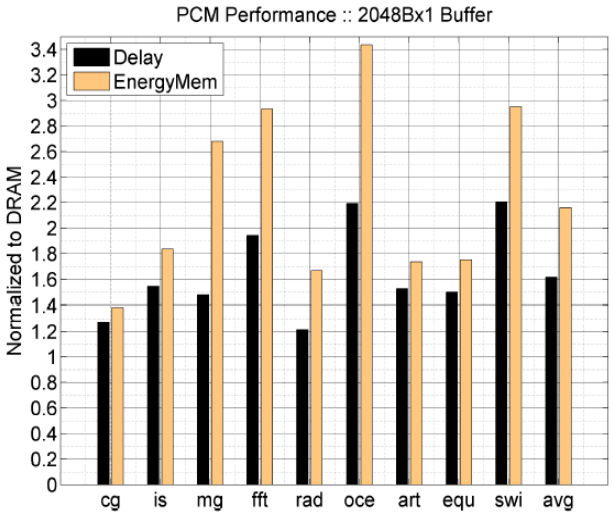


Figura 4.7. [1] Ritardi e costi energetici mostrati da un'applicazione utilizzante dispositivi PCM con architettura [1X2048B], in luogo di una memoria standard DRAM (rispetto ai valori evidenziati da quest'ultimo dispositivo è effettuata la normalizzazione).

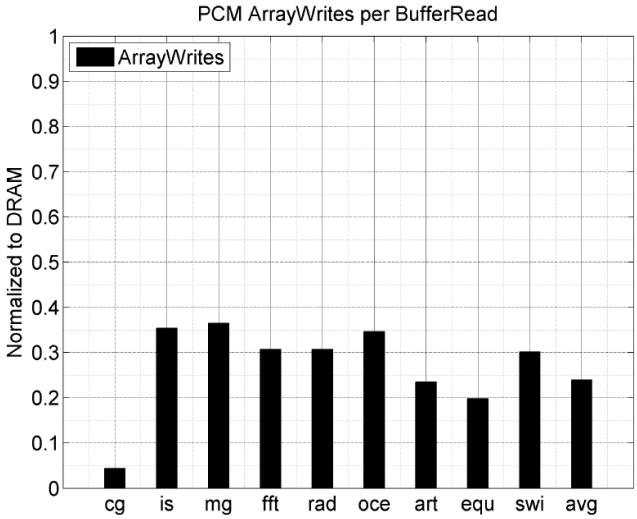


Figura 4.8. [1] Percentuali delle operazioni di lettura che richiedono l'esecuzione di nuova scrittura, a causa della presenza nel buffer interno di dati spuri (appunto rilevati durante il readout del medesimo registro).

Il primo dei due grafici precedenti mostra i ritardi ed i costi energetici riscontrabili durante l'esecuzione di un'applicazione che utilizza, in luogo di una comune memoria DRAM, un dispositivo PCM avente architettura ad

unico buffer interno da 2048B.

In particolare, dall'analisi della Fig. 4.7 emerge che i ritardi rilevati durante l'esecuzione dell'applicazione sono al minimo pari a 1.2 volte (vedi PCM radix), fino a un massimo di 2.2 volte (vedi PCM ocean o swim) quelli evidenziati dalla stessa applicazione nel caso in cui la medesima avesse a propria disposizione una memoria DRAM standard.

Per quanto riguarda invece i consumi energetici, sempre la Fig. 4.7 mette in evidenza come anche i valori di queste grandezze crescono da 1.4 volte (vedi PCM cg) fino ad un massimo di 4.4 volte (vedi PCM ocean) rispetto ai dispendi energetici mostrati nel caso di utilizzo di una DRAM da parte della stessa applicazione.

In realtà, sia i ritardi, sia i costi energetici individuati, tendono a sovrastimare quelli reali dato che anche in situazioni in cui il dispositivo PCM sia utilizzato in modo intensivo, bisogna comunque tener presente che parte dell'energia e dei tempi di latenza associati all'esecuzione di un'applicazione, sono attribuibili sì agli accessi in memoria, ma anche alla computazione dello stesso programma.

Inoltre, è importante non dimenticare che spesso le scritture di celle appartenenti ad una PCM, caratterizzate per l'appunto da alti costi energetici ed elevati tempi di latenza, sono in realtà molto meno frequenti dei processi di scrittura che avvengono comunemente in una DRAM (nota: le letture di una PCM non sono distruttive, e non necessitano quindi di successive operazioni di scrittura come appunto accade invece per le operazioni di lettura di un qualsiasi dispositivo DRAM. Questo aspetto ovviamente incide “positivamente“ sui costi energetici e sui ritardi effettivamente associabili all'esecuzione di un'applicazione che abbia a propria disposizione una PCM, ma ciò non è mostrato nella Fig. 4.7).

Per quanto riguarda invece la Fig. 4.8, essa mette in rilievo come una

percentuale in media pari al 28% delle operazioni di lettura delle PCM, richieda in realtà una nuova operazione di scrittura della riga di celle a causa del fatto che valori logici spuri sono stati individuati durante il readout del buffer interno.
Alla luce di quanto detto precedentemente, è evidente che i dati emersi dai due grafici di sopra evidenziano l'esistenza di differenze non trascurabili tra dispositivi PCM e DRAM, pertanto è necessario colmare le stesse se si intende adoperare memorie PCM come cache di più basso livello all'interno di un sistema a microprocessore.
A questo proposito, il gap di performance presente tra PCM e DRAM può essere ragionevolmente livellato mediante alcune soluzioni circuitali.

4.10 – Nuova organizzazione dei buffer interni di una memoria PCM.

La definizione di una nuova organizzazione per i buffer interni di un dispositivo PCM consente in genere di nascondere e ridurre rispettivamente i tempi di latenza ed i costi energetici delle operazioni di lettura/scrittura, rendendo così queste memorie delle valide alternative alle DRAM.
La riorganizzazione dell'area interna delle PCM prevede prima di tutto la riduzione della lunghezza dei buffer interni ed un contestuale aumento del numero di righe dedicate per l'alloggiamento degli stessi buffer. Questi ripensamenti organizzativi producono effetti positivi anche in termini di riduzione di area interna visto che, con la diminuzione della grandezza dei singoli buffer, si ha una corrispondente diminuzione del numero di s.a. da includere all'interno del dispositivo. L'area così “recuperata” è poi di solito destinata ad alloggiare righe multiple di latch.
Con questa nuova organizzazione, i dispositivi PCM evidenziano in genere

una riduzione dei costi energetici di scrittura, anche se al contempo fanno registrare un degrado delle performance relative all'esecuzione di applicazioni (ad esempio, diminuiscono le possibilità di aggregare ed eseguire insieme più operazioni di scrittura).

4.11 – Analisi dell'occupazione e della densità d'area che caratterizza la nuova organizzazione interna.

		PCM	DRAM
Array			
A	bank size (MB)	16	16
C	cell size (F^2)	9MLC, 12MLC	6
Periphery			
S	sense amplifer (T @ $250\lambda^2$/T)	44	14
	sense amplifer (F^2)	2750	875
L	latch (T @ $250\lambda^2$/T)	8	0
	latch (F^2)	500	0
D	decode 2-AND (T @ $1000\lambda^2$/T)	6	0
	decode 2-AND (F^2)	250	0
Buffer Organization			
W	buffer width (B)	64::2x::2048	2048
R	buffer rows (ea)	1::2x::32	1

Tabella 4.9. [1] Tabella dei parametri d'area espressi secondo le unità di misura: T (numero di transistor che occupano equivalentemente l'area considerata), F² (square feature sizes) e λ (layout design). Le ultime due righe della tabella evidenziano inoltre la riorganizzazione subita dai buffer interni del dispositivo PCM, e cioè si suppone che lo stesso dispositivo di memoria disponga di 32 registri da 64B ciascuno (al contrario invece della DRAM a confronto, che possiede architettura dei buffer interni [1X2048B]).
Nota: le aree espresse numero di transistor T per i Sense Amplifiers, sono state determinata da Sinha et al. [9], mentre i valori di densità λ²/T sono stati ricavati dalla Tabella 1.10 di Weste e Harris [10].

La tabella precedente mostra in termini quantitativi la riorganizzazione dei circuiti periferici delle bitline (vedi s.a., latch e nuovi circuiti di decodifica), nonché dei buffer interni (vedi circuiti di decodifica per le righe dedicate all'allocazione ulteriore dei registri), apportata ad un dispositivo PCM da 16MB utilizzato come proprio riferimento.

I componenti circuitali prima elencati sono considerati come elementi che hanno un alto impatto sul risparmio di occupazione e sull'incremento di densità d'area che si ottiene in conseguenza della riorganizzazione interna della memoria PCM.

La ragione per cui i parametri mostrati nella tabella di sopra sono espressi secondo le unità di misura T, F^2 e λ, sta nel fatto che in questo modo l'analisi che si ottiene è indipendente dal processo tecnologico di realizzazione del dispositivo di memoria. A tal proposito è necessario fare alcune precisazioni, e cioè: con *F* (feature size) è indicata l'unità di misura del processo produttivo con cui è realizzata la memoria che si considera (ad esempio 90nm), nonché è possibile quindi esprimere in termini di F^2 (square feature sizes) l'area occupata da una cella o da un qualsiasi dispositivo interno di memoria.

In maniera del tutto equivalente, la stessa superficie ingombrata da una cella o da un componente interno, può essere espressa in termini di numero di transistor (T) che per l'appunto occuperebbero la stessa estensione d'area dell'elemento circuitale considerato. Un'ultima precisazione riguarda invece il layout design λ e l'unità di misura della densità d'area λ^2/T, e cioè: la prima di queste due grandezze è in relazione con F secondo l'equazione $\lambda = 2F$; mentre la seconda grandezza consente di esprimere una qualsiasi area occupata da un componente circuitale in termini di layout design quadratico λ^2, se appunto si conosce il valore di λ^2/T e di T del dispositivo in esame.

Alla luce di quanto detto di sopra, è possibile quindi realizzare una prima distinzione tra componenti interni con alta densità d'integrazione, caratterizzati per l'appunto da valori di densità dell'ordine di $250\lambda^2/T$ (in questo caso si parla di dense datapath circuits), ed elementi circuitali con bassa densità, con valori pari 1000 λ^2/T per lo stesso parametro

precedente (essi sono noti come controll circuits*)* [10].

4.12 – Analisi del modello d'occupazione d'area mostrato nella Tabella 4.9 per un generico dispositivo PCM.

Generalmente, l'occupazione d'area da parte di una singola cella di PCM oscilla fra i $6F^2$ ed i $20F^2$, a differenza appunto di quanto accade per una cella di DRAM, che ingombra un'area pari a circa $6F^2$ in virtù del suo design e della maggiore esperienza maturata nel relativo campo implementativo.

La differenza esistente fra le celle DRAM e quelle PCM è in primo luogo determinata dal dispositivo di accesso utilizzato dalle stesse, e cioè: se nel primo caso è in genere adoperato un relativamente piccolo MOSFET, nel caso invece di celle PCM è preferito l'utilizzo di BJT, i quali appunto garantiscono tempi d'accesso più contenuti, ma che tuttavia implicano un'occupazione d'area maggiore rispetto a quella evidenziata dagli altri elementi di accesso.

Alla luce di quanto detto, si dimostra che celle di PCM utilizzanti BJT presentano conseguentemente un'occupazione d'area pari a $9\div12F^2$, nonché per fare in modo che i dispositivi PCM offrano densità d'integrazione uguali a quelle di DRAM con celle da $6F^2$, le memoria PCM sono realizzate in modo da disporre di celle Multi-Level.

In conseguenza di ciò, anche le memorie a cambiamento di fase sono in grado di offrire densità di $4.5\div6F^2$/bit.

4.13 – Analisi dei costi energetici e dei ritardi mostrati dai dispositivi PCM con nuova organizzazione interna.

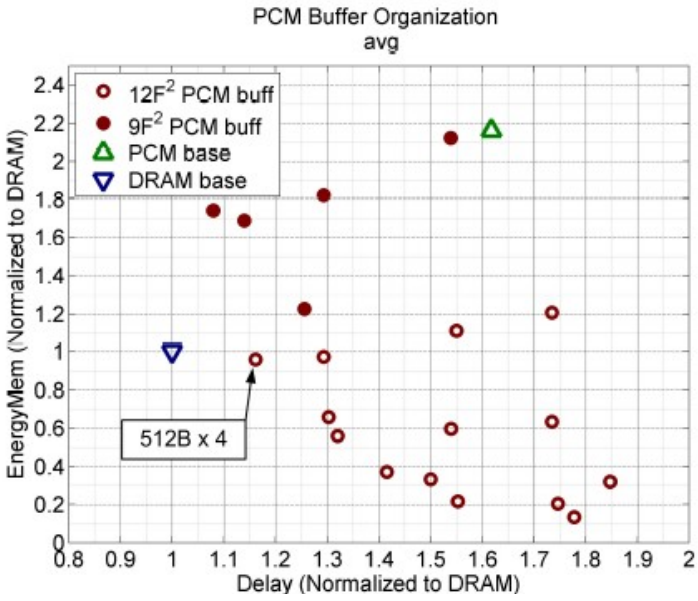


Figura 4.10. [1] Grafico dei costi energetici e dei ritardi (normalizzati rispetto a quelli di una DRAM standard) evidenziati da dispositivi di memoria PCM con celle 9-12F².

Nel grafico di sopra sono rappresentati: (1) mediante dei triangoli, le prestazioni (in termini di ritardi e costi energetici) di un comune dispositivo DRAM e di uno PCM, entrambi con architettura interna a buffer unico da 2048B (questi chip sono il riferimento per l'analisi condotta); (2) per mezzo di cerchi vuoti e pieni, i ritardi e i dispendi energetici associabili in media rispettivamente a dispositivi PCM aventi celle $12F^2$ e $9F^2$, che hanno subito una riorganizzazione interna dei propri dispositivi periferici e dei propri buffer.

Detto ciò, il grafico di Fig. 4.10 evidenzia come, sebbene i dispositivi PCM con celle $9F^2$ presentino una riduzione più consistente dei ritardi rispetto a quella evidenziata dai chip di memoria a cambiamento di fase con celle $12F^2$, proprio quest'ultima categoria di PCM mostra però costi energetici praticamente identici a quelli della DRAM di riferimento, ed addirittura in

alcuni casi anche inferiori rispetto a quelli evidenziati dalla stessa DRAM.
L'analisi condotta dagli autori dell'articolo di riferimento implica che l'ottimizzazione dei ritardi e dei costi è raggiunta in corrispondenza di un'architettura interna con quattro buffer da 512B (4X512B), utilizzata però da una matrice di memoria con celle $12F^2$.
Inoltre, sempre supponendo di voler ulteriormente minimizzare i consumi energetici dei dispositivi PCM, è possibile dimostrare come in realtà il dispendio energetico tende ad aumentare leggermente al crescere del numero di righe dedicate ai buffer interni. Questo fenomeno è determinato dal fatto che un dispositivo PCM con buffer disposti su più righe ha una maggiore richiesta di potenza di background.
Tuttavia, la riduzione del dispendio di "energia dinamica" ottenuta con l'aumento del numero di righe di buffer (e quindi dovute all'ottimizzazione delle procedure di lettura/scrittura), è molto più netta rispetto all'aumento del consumo di Background Energy, pertanto sono sempre preferibili dispositivi aventi architetture con più buffer interni.

4.14 – Valutazione del processo di scaling delle memorie PCM e confronto con quello delle DRAM.

Il processo di scaling delle DRAM è assai complesso quanto problematico, dato che la miniaturizzazione dei capacitori e dei transistor di ogni singola cella in realtà presentano numerosi inconvenienti, e cioè: (1) la riduzione delle dimensioni circuitali deve avvenire in modo compatibile con i meccanismi di conservazione della carica elettrica da parte dei condensatori presenti nelle celle (di qui i problemi legati alla miniaturizzazione dei capacitori); (2) il processo di scaling determina una riduzione delle dimensioni dei transistor di accesso, a cui appunto

corrisponde generalmente un aumento delle correnti di perdita della singola cella, e di conseguenza un degrado dei tempi di ritenzione dei dati salvati (nota: una riduzione dei tempi di ritenzione comporta a sua volta una maggiore richiesta energetica durante l'esecuzione delle operazioni di refresh).

Questi due aspetti incidono fortemente sull'affidabilità e sui i costi energetici dei dispositivi DRAM, nonché l'ITRS stima che realizzare memorie DRAM andando al di sotto della soglia rappresentata da un processo produttivo a 40nm è molto difficile, quanto improbabile [5].

Al contrario, lo stesso ITRS sostiene che per le memorie PCM sia possibile raggiungere addirittura processi produttivi di 9nm [5].

Il processo di scaling, inoltre, presenta vantaggi energetici più consistenti per i dispositivi PCM visto che, nel realizzare uno scaling del processo produttivo [80nm → 40nm], l'ITRS ipotizza che la riduzione dei costi energetici che ne deriverebbe in corrispondenza dei 40 nm sarebbe pari a: 2.4volte quelli dei dispositivi PCM prodotti con 80nm; e 1.5volte quelli delle memorie DRAM realizzate con processo 80nm [5].

A tal proposito, la Fig. 4.11 è assai esplicativa della convenienza dei dispositivi PCM implementati con processo 40nm, rispetto ad un chip standard DRAM, appunto anch'esso realizzato con processo produttivo 40nm. Dal grafico di Fig. 4.11 è possibile notare come i costi energetici associati alle memoria PCM siano pari in media al 61% di quelli attribuibili al dispositivo DRAM di riferimento.

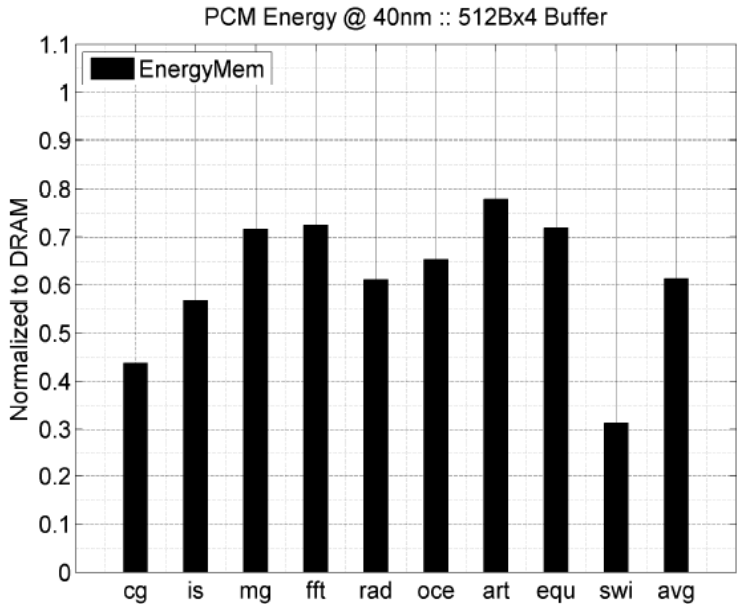


Figura 4.11. [1] Riduzioni dei costi energetici mostrate dai dispositivi PCM di benchmarking realizzati con processo 40nm e con architettura [4X512B] (per i buffer interni), rispetto ad una DRAM implementata anch'essa con processo a 40nm. Si tenga presente che questa analisi comparativa non tiene conto dell'incremento dei costi energetici dovuti alle usuali operazioni di refresh del dispositivo di DRAM.

Capitolo 5 – Confronto tra le tecnologie analizzate e conclusioni.

Dopo aver analizzato nei capitoli 2, 3 e 4, quali sono le più moderne e promettenti tecnologie implementative di memorie non volatili, quest'ultimo capitolo è dedicato alla realizzazione di un confronto tra le tecnologie precedentemente illustrate e quelle che ad oggi sono le più impiegate nell'ambito della fabbricazione di componenti di memoria per sistemi digitali.

		DRAM		SRAM [2]	Floating-Gate [3]		FeRAM	MRAM	PCM
		Stand-Alone [1]	Embedded [2]		NOR	NAND			
Architettura di cella		1T1C	1T1C	6T	1T	1T	1T	1(2)T1C	1T1R
Feature size F (nm)	2007	68	90	65	90	90	180	90	65
	2022	12	25	13	18	18	65	22	18
Area di Cella (F^2)	2007	6	12	140	10	5	22	20	4,8
	2022	6	12	140	10	5	12	16	4,7
Tempo Accesso in Lettura	2007	< 10 ns	1 ns	0,3 ns	10 ns	50 ns	45 ns [4]	20 ns [7]	60 ns [9]
	2022	< 10 ns	0.2 ns	70 ps	2 ns	10 ns	< 20 ns [5]	< 0,5 ns	< 60 ns
Tempo Accesso in Scrittura	2007	< 10 ns	0,7 ns	0,3 ns	1 s	1 ms	10 ns [6]	20 ns [7]	50÷120 ns [9]
	2022	< 10 ns	0,2 ns	70 ps	1 s	1 ms	1 ns [J]	< 0,5 ns [8]	< 50 ns
Tempo Fase Cancellazione	2007	[F]	[F]	[F]	10 ms	0,1 ms	[F]	[F]	[F]
	2022	[F]	[F]	[F]	10 ms	0,1 ms	[F]	[F]	[F]
Tempo Ritenzione Dati	2007	64 ms	64 ms	[B]	> 10 anni	> 10 anni	> 10 anni	> 10 anni	> 10 anni
	2022	64 ms	64 ms	[B]	> 10 anni	> 10 anni	> 10 anni	> 10 anni	> 10 anni
Endurance	2007	$> 3X10^{16}$	$> 3X10^{16}$	$> 3X10^{16}$	$> 10^5$	$> 10^5$	10^{14}	$> 3X10^{16}$	10^8
	2022	$> 3X10^{16}$	$> 3X10^{16}$	$> 3X10^{16}$	$> 10^5$	$> 10^5$	$> 10^{16}$	$> 10^{16}$	10^{15}
Energia Fase di Scrittura (J/bit)	2007	$5 X 10^{-15}$ [A]	$5 X 10^{-15}$	$7 X 10^{-16}$	$> 10^{-14}$ [C]	$> 10^{-14}$ [C]	$3 X 10^{-14}$ [D]	$7 X 10^{-11}$ [1]	$5 X 10^{-12}$ [E]
	2022	$2 X 10^{-15}$ [A]	$2 X 10^{-15}$	$2 X 10^{-17}$	$> 10^{-15}$ [C]	$> 10^{-15}$ [C]	$5 X 10^{-15}$ [D]	$2 X 10^{-11}$ [1]	$< 10^{-13}$ [E]

Note: [A] Energia stimata come $0,5*C*V^2$, con C = 25fF e V = 0,65V;0,35V per il 2007 ed il 2022;
[B] Le SRAM conservano i dati sino a che sono alimentate, nonché non hanno bisogno di operazioni di refreshing come nel caso delle DRAM;
[C] E' proprio il limite inferiore che caratterizza l'"Effetto Fowler-Nordheim" utilizzato per le fasi di scrittura/cancellazione;
[D] Energia stimata come $0,5*\rho*A*V$, con ρ = 10,9μC/cm^2;30μC/cm^2 e V = 1,5V;0,7V e A = 0,33μm^2;0,069μm^2 rispettivamente per il 2007 ed il 2022;
[E] Energia stimata come $0,5*I^2*R*t_w$, con I = 235μA;13μA e R = 3,54kΩ;35,4kΩ e t_w = 50ns;<50ns rispettivamente per il 2007 ed il 2022;
[F] Non sono necessarie operazioni di cancellazione per le DRAM, SRAM, FeRAM, MRAM e PCM;

Tabella 5.1. [1] Confronto tra parametri tipici di dispositivi sviluppati con tecnologia magnetoresistiva, ferroelettrica ed a cambiamento di fase ed i parametri caratteristici dei componenti di memoria SRAM, DRAM, Flash (precisazioni: anche le note che compaiono al di sotto di questa tabella sono state tratte da [1]).

La Tab. 5.1 mostra in sequenza: i processi produttivi, le aree occupate dalle singole celle, i tempi necessari per eseguire un'operazione di lettura/scrittura/cancellazione (quest'ultima, qualora richiesta), la durata dell'intervallo temporale di ritenzione dei dati, il grado di endurance e l'energia richiesta per eseguire la scrittura di un bit, che caratterizzano tipicamente sia i dispositivi DRAM, SRAM e le Memorie Flash (con architettura a NOR e a NAND), sia i più moderni componenti di memoria FeRAM, MRAM e PCM.

Partendo dal parametro di area occupata dalla singola cella, i dispositivi che risultano essere quelli più svantaggiosi sono le SRAM, le quali appunto presentano per tale parametro un valore di $140F^2$ (si stima tra l'altro che tale valore non possa decrescere in futuro, vedi “Previsione 2022” in Tab. 5.1 e Grafico 5.2), pari appunto a 5 volte il valore medio di occupazione di area associabile alle celle degli altri dispositivi esaminati (la media aritmetica dei valori mostrati in tabella per l'anno 2007, è pari a $27.475F^2$).

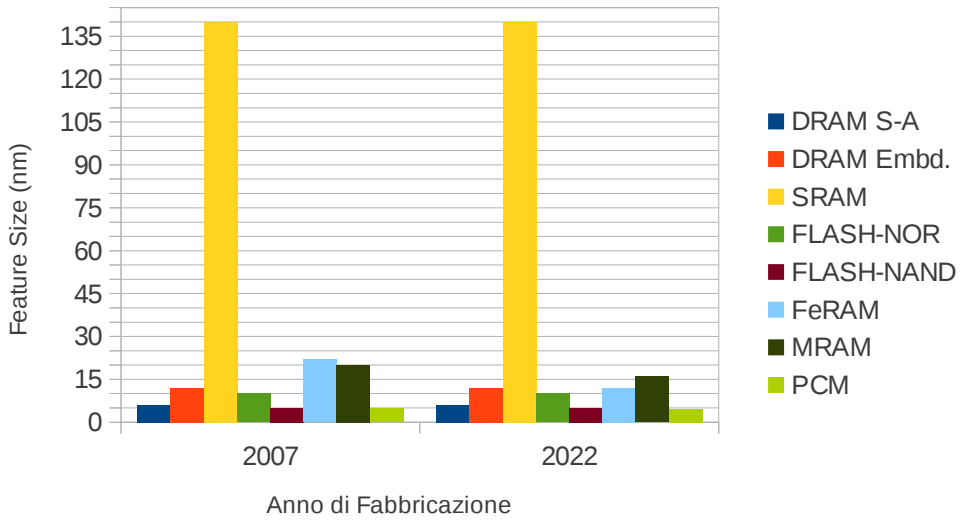


Grafico 5.2. Confronto tra l'area occupata dalle singole celle dei dispositivi di memoria prodotti nel 2007, e quella che si stima che abbiano le stesse locazioni di memoria nel 2022.

Per questo motivo, quindi, l'implementazione di una SRAM anche attraverso il più semplice *processo di fabbricazione*, implica costi molto più elevati rispetto a quelli generalmente sostenuti per la realizzazione di altre tipologie di memoria, nonché l'utilizzo delle stesse SRAM come componenti interni di Sistemi Embedded che richiedono grandi capacità dati è fortemente sconveniente, dato che da questo punto di vista (e cioè relativamente alla capacità dati e/o alla densità d'integrazione) risultano essere più vantaggiose altre famiglie di dispositivi (l'"unica" convenienza nell'adozione di dispositivi SRAM sta nell'elevate velocità di lettura/scrittura dati offerte da questi componenti).

Sempre per quanto riguarda l'occupazione di area da parte delle celle di memoria, i dispositivi PCM sono quelli sicuramente più convenienti in questo ambito (vedi valore di $4.8F^2$), nonché l'adozione di questa categoria di componenti per esempio per la sostituzione di elementi DRAM potrebbe risultare particolarmente vantaggiosa, poiché la *non-volatilità* delle memorie a cambiamento di fase consentirebbe di eliminare i fondamentali cicli di refresh proprie delle celle dinamiche di DRAM, che rende quest'ultima categoria di elementi di memoria per l'appunto maggiormente dispendiose dal punto di vista energetico (quest'ultimo aspetto limita infatti l'impiego di componenti di memoria DRAM all'interno di dispositivi portatili, dato che l'energia richiesta dai cicli, incide negativamente sulla durata delle batterie di tali sistemi digitali).

In realtà, un'ulteriore valida idea per quanto riguarda il contesto appena discusso, potrebbe essere quella di adoperare componenti FeRAM in luogo di DRAM dato che, nonostante presentino un livello di densità d'integrazione più basso rispetto a quello delle memorie a cambiamento di fase, esse evidenziano altresì consumi energetici pressoché simili a quelli delle DRAM (6 volte superiori rispetto a quelli richiesti dalle DRAM), e ciò

sarebbe ancor più vantaggioso nell'ottica della minimizzazione dei consumi di potenza, visto che al contrario i dispositivi PCM richiedono energie nettamente maggiori per eseguire le operazioni di scrittura (sono 1000 volte superiori rispetto a quelle richieste dalle DRAM).

Passando ora ai dispositivi di memoria non volatili attualmente diffusi, i dati relativi alle Flash Memory mettono in luce come questi componenti presentino sia ottimi livelli di densità d'integrazione, sia bassi tempi di accesso in lettura, per l'appunto nettamente migliori rispetto ai valori mostrati per questi stessi parametri dai componenti FeRAM, MRAM e PCM.

Tuttavia, gli svantaggi legati alla tecnologia Flash sono rappresentati principalmente dagli elevatissimi tempi di scrittura dei valori logici e dalle alte tensioni richieste per eseguire la programmazione delle celle (quest'ultime limitano fortemente infatti il n° di cicli di scrittura eseguibili sulle celle Flash).

Inoltre le Memorie Flash, a differenza dei componenti FeRAM, MRAM e PCM, eseguono preliminarmente operazioni di cancellazione dei valori logici presenti nelle celle, ogni qualvolta appunto sia richiesta una ri-programmazione dello stato di bit delle celle.

Ciò ovviamente incide negativamente sui tempi di latenza complessivamente percepiti dall'esterno e associabili all'esecuzione di un'operazione di scrittura/riprogrammazione per una cella di Flash Memory, nonché favorisce l'utilizzo di FeRAM, MRAM e PCM.

A questo punto, volendo approfondire ulteriormente l'analisi dei parametri mostrati in tabella con riferimento però alle sole tecnologie discusse nei capitoli 2, 3 e 4, prendendo nuovamente in considerazione il grafico 5.2, si capisce come un nuovo processo di miniaturizzazione sia considerato possibile solo per le celle di FeRAM e di MRAM, anche se si prevede che

lo scaling applicato a queste tecnologie non consentirà comunque di raggiungere l'alta densità d'integrazione tipico delle PCM.

In virtù quindi della ridottissima area di cella tipica delle memorie a cambiamento di fase, sono stati già implementati dispositivi e prototipi PCM con capacità dati che raggiungono anche gli 8 Gbits (vedi prototipo del Febbraio 2012 prodotto da Samsung con processo a 20nm [11]), a differenza invece dei dispositivi MRAM e FeRAM che presentano al più capacità dati rispettivamente di 32Mbit (vedi prototipo MRAM del Giugno 2009 realizzato da Hitachi, in collaborazione con la “Tohoku University” [12]) e di 4Mbits circa [11].

Siccome, i dispositivi di memoria con maggiore grado di densità d'integrazione sono anche i componenti che permettono di ottenere un minor costo di fabbricazione per singola cella, allora si presume che le PCM siano anche gli elementi di memoria che nel tempo diventeranno quelli più economici (vedi anche la possibilità di implementare celle Multi-Level).

In virtù di quanto detto di sopra, le memorie a cambiamento di fase rappresentano difatti la famiglia di dispositivi che sostituiranno con maggiore probabilità le memorie Flash con architettura NOR [11] (i dispositivi Flash NAND presentano infatti capacità dati ampiamente più elevate rispetto alle Flash NOR e alle PCM [11]).

In realtà, il contesto principale nel quale le PCM evidenziano al meglio le loro potenzialità, è quello dell'industria militare ed aerospaziale. In particolare, siccome l'utilizzo dei classici dispositivi di memorizzazione non volatili è assolutamente inadatto a contesti in cui i raggi cosmici possono molto facilmente mettere a serio repentaglio l'affidabilità dei dati salvati [11], le PCM invec, utilizzando processi di scrittura dati che sfruttano un fenomeno fisico ben diverso da quello dell'iniezione di elettroni tipico delle

memorie a floating-gate, presentano conseguentemente elevati gradi di tolleranza agli effetti per l'appunto provocati dall'esposizione ai raggi spaziali.

In realtà, esistono anche “aspetti negativi” legati all'utilizzo di questi componenti di memoria, e cioè basti pensare infatti ai più elevati tempi di lettura/scrittura messi in evidenza appunto dai dispositivi a cambiamento di fase rispetto alle memorie magnetoresistive ed a quelle ferroelettriche (vedi Grafico 5.3).

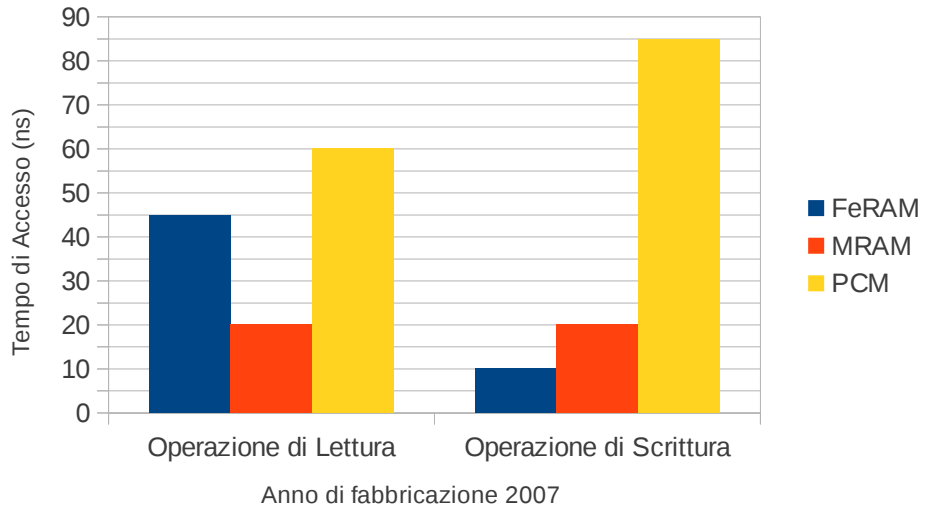


Grafico 5.3. Tempi di accesso in lettura ed in scrittura delle memorie realizzate con tecnologia magnetoresistiva, ferroelettrica ed cambiamento di fase (nota: per le PCM si è rappresentato come tempo di accesso in scrittura un valore pari a 85ns, e cioè il valore medio dei 50ns e 120ns utilizzati per salvare un valore logico rispettivamente alto/basso).

Il grafico di sopra, indica inoltre quale sia la categoria di memorie che meglio riesce a minimizzare sia i tempi richiesti per eseguire un'operazione di lettura, sia di quelli tipici di una fase di scrittura, ovvero i dispositivi MRAM.

Lo svantaggio implicitamente connesso all'utilizzo dei componenti di memoria magnetoresistivi riguarda i costi energetici delle operazioni di scrittura, dato che sono 2333,33 volte superiori rispetto a quelli evidenziati

dalle FeRAM, e 14 volte maggiori rispetto all'energia tipicamente richiesta per la scrittura di un bit all'interno di un dispositivi PCM.

Riassumendo il tutto, dunque, si può notare come: (1) la tecnologia a cambiamento di fase rappresenta quella migliore per la realizzazione di dispositivi dalle grandi capacità dati e dai contenuti costi di fabbricazione per singola cella di memoria; (2) la tecnologia magnetoresistiva consente invece di implementare dispositivi veloci in fase di scrittura/lettura; (3) la tecnologia ferroelettrica è infine quella da adottare se si il proprio obiettivo primario è la fabbricazione di elementi di memoria con minimo dispendio energetico.

Sia le MRAM, che le FeRAM trovano largo utilizzo all'interno dei System-On-Chip utilizzati da più o meno complesse applicazioni embedded, nonché si contendono le medesime fette di mercato (di nicchia) lasciate libere dalle Flash NAND, che appunto tuttora rappresentano la tecnologia principalmente utilizzata per l'implementazione di memorie non volatili.

In particolare, i nuovi dispositivi magnetoresistivi e ferroelettrici sono utilizzati principalmente:

- all'interno di notebook, personal computer, ebook, camere digitali, telefoni cellulari, apparecchiature biomediche ed aerospaziali (vedi scatole nere), ecc. [12];
- all'interno dei comuni contatori di energia elettrica, microcontrollori industriali, dispositivi con RFID, apparecchiature di medicina, ecc. [13].

Per capire meglio quanto precedentemente scritto, il presente lavoro di tesi si conclude con la realizzazione di un confronto delle prestazioni e delle caratteristiche tecniche proprie di tre componenti di memoria attualmente presenti in commercio.

A tal proposito, la Tab. 5.4 mostra i valori dei parametri caratteristici dei dispositivi: (1) FeRAM “FM20L08 – 1Mbit Bytewide FRAM Memory Extended Temp.”, prodotto da Ramtrom; (2) MRAM “MR2A16A ”, realizzato da Freescale Semiconductor e Motorola; (3) PCM “P8P Parallel Phase Change Memory (PCM)” commercializzato da Micron Technology (nota: i valori sono stati ricavati analizzando i datasheet [14], [15] e [16]).

	FeRAM	MRAM	PCM
Casa Produttrice	Ramtrom	Freescale Semiconductor & Motorola	Micron Technology
Modello	FM20L08	MR2A16A	P8P Parallel Phase Change Memory
Capacità Dati	1Mbit	4Mbit	128Mbit
Capacità (n° singoli bit)	1048576	4194304	134217728
Architettura Interna	128K x 8bit	256K x 16bit	127Blocchi x 128KB + 4Blocchi x 32KB
Modalità Memorizzazione	non volatile	non volatile	non volatile
Tensioni di Lavoro (V)	3,3	3,3	2,7 ÷ 3,3
Corrente di Lavoro (mA)	22	55 ÷ 105 *3	15 ÷ 35 *6
Corrente di Standby (µA)	20	18	80
Tempo Accesso in Lettura (ns)	≥ 350 (≥ 65) *1	≥ 35 (≤ 35) *4	≥ 115 (≤ 115) *7
Tempo Accesso in Scrittura (ns)	≥ 350 (≥ 65) *2	≥ 35 (≥ 18) *5	≥ 50 *8
Modalità Scrittura Dati	dati sovrascritti	dati sovrascritti	dati sovrascritti
Cancellazione Dati	non necessaria	non necessaria	non necessaria
Endurance (n° cicli scrittura)	illimitata	non disponibile	10^6
Ritenzione dati (anni)	> 10	> 20	10
Prezzo dispositivo (€)	22,95	19,49	non disponibile
Costo per cella (€ x 1cella) *9	2,19E-05	4,65E-06	non disponibile

Note: *1 → E' mostrata la durata temporale dell'intero ciclo di lettura, assieme al reale tempo di accesso in lettura al dato (vedi valore in parentesi, è stato determinato come somma di t_{AS}+t_{CE}, supponendo di realizzare una lettura controllata con il pin $\overline{CE}$).

*2 → E' mostrata la durata temporale dell'intero ciclo di scrittura, assieme al reale tempo di accesso in scrittura del dato (vedi valore in parentesi, è stato determinato come somma di t_{AS}+t_{CW}, supponendo di realizzare una scrittura controllata con i pin $\overline{CE}$ e $\overline{WE}$).

*3 → il primo valore indicato è l'intensità di corrente di lavoro per le Fasi di Lettura, mentre il secondo valore rappresenta l'intensità di corrente di lavoro per le Fasi di Scrittura.

*4 → E' mostrata la durata temporale dell'intero ciclo di lettura, assieme al reale tempo di accesso in lettura al dato (vedi t_{AVAV} e t_{AVQV} per la lettura in modalità 2).

*5 → E' mostrata la durata temporale dell'intero ciclo di scrittura, assieme al reale tempo di accesso in scrittura del dato (vedi valore del parametro t_{AVWH}).

*6 → I valori di intensità di corrente mostrati, sono riferiti rispettivamente all'esecuzione di un'operazione di lettura ed a una di scrittura.

*7 → E' mostrata la durata temporale dell'intero ciclo di lettura, assieme al reale tempo di accesso in lettura al dato (vedi t_{AVAV} e t_{AVQV} dei diagrammi temporali).

*8 → E' mostrato il tempo di accesso in scrittura di un dato (vedi t_{AVWH}).

*9 → I valori di questo parametro sono stati ricavati dividendo il costo del dispositivo per il numero di bit di capacità.

Tabella 5.4. Confronto tra i parametri caratteristici di dispositivi FeRAM, MRAM e PCM attualmente in commercio. Precisazioni: i dati che sono riportati in tabella e le spiegazioni delle note (esclusa la n° 9) sono sta ricavate dai datasheet [14], [15] e [16] dei componenti analizzati.

I dati mostrati in tabella confermano effettivamente l'analisi condotta precedentemente sulle nuove tecnologie magnetoresistive, ferroelettriche

ed cambiamento di fase circa il consumo di potenza (vedi voci “Tensione di Lavoro” e “Corrente di Lavoro”), la velocità di lettura e scrittura (vedi voci “Tempo di accesso in Lettura/Scrittura”) e la possibilità d'avere dispositivi con capacità dati più o meno estese (vedi voce “Capacità Dati”).

Dalla tabella di sopra è possibile in ultimo notare come il dispositivo FeRAM considerato sia in realtà più costoso di quello MRAM preso in esame, nonostante il medesimo abbia una capacità dati inferiore rispetto per l'appunto a quella del componente magnetoresistivo.

Infine, sempre per quanto riguarda i costi economici di fabbricazione degli elementi di memoria, si noti come l'ultima riga della tabella si traduce nel seguente grafico:

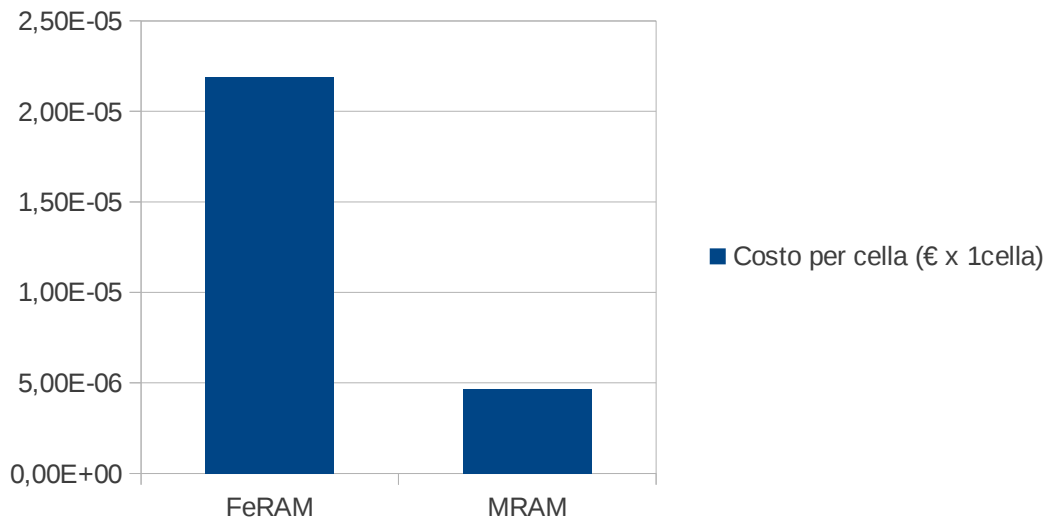
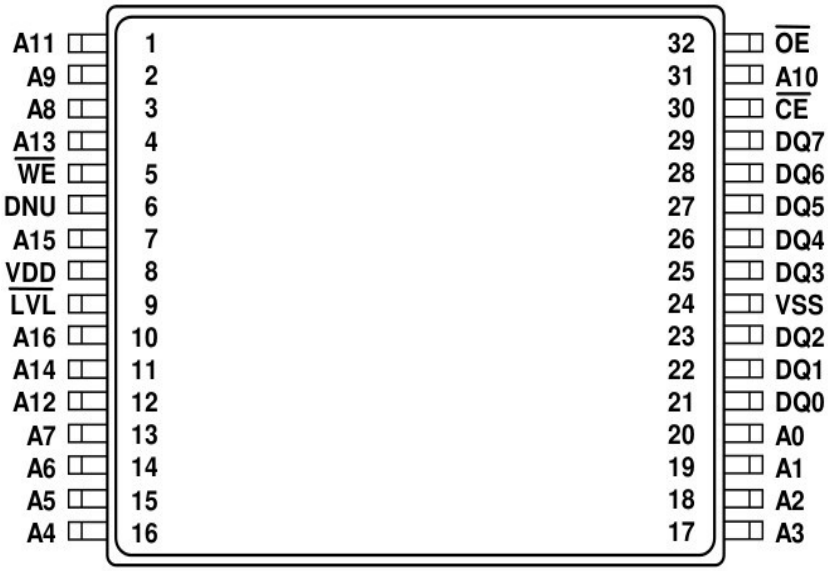


Grafico 5.5. Confronto tra il Costo Per Cella del dispositivo “FeRAM FM20L08 Ramtrom” e quello del componente “MRAM MR2A16A Freescale Semiconductor&Motorola”.

che per l'appunto mette in evidenza come il costo di una singola cella della memoria ferroelettrica Ramtrom in esame sia in realtà 4.71 volte quello di una cella appartenente alla memoria magnetoresistiva Freescale-Motorola considerata.

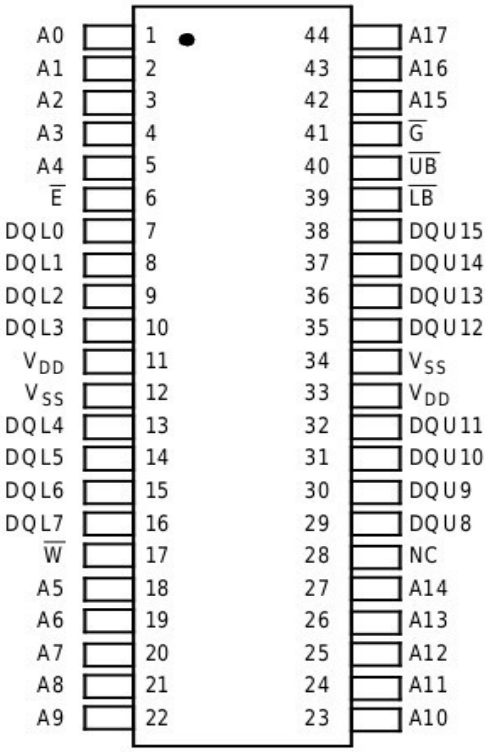
5.1 – Approfondimento Capitolo 5: Pinouts dei dispositivi MRAM, FeRAM e PCM analizzati in Tabella 5.4.



Nome Pin	Tipo	Funzione
A (16:0)	Input	Pin degli Indirizzi
$\overline{CE}$	Input	Chip Enable
$\overline{OE}$	Input	Output Enable
$\overline{WE}$	Input	Write Enable
$\overline{LVL}$	Output	Level Voltage Lockout (*1)
DQ (7:0)	Input/Output	Dati I/O
V_{DD}	Input	Tensione di Alimentazione
V_{SS}	Input	Massa
DNU	---	Pin da Non Collegare

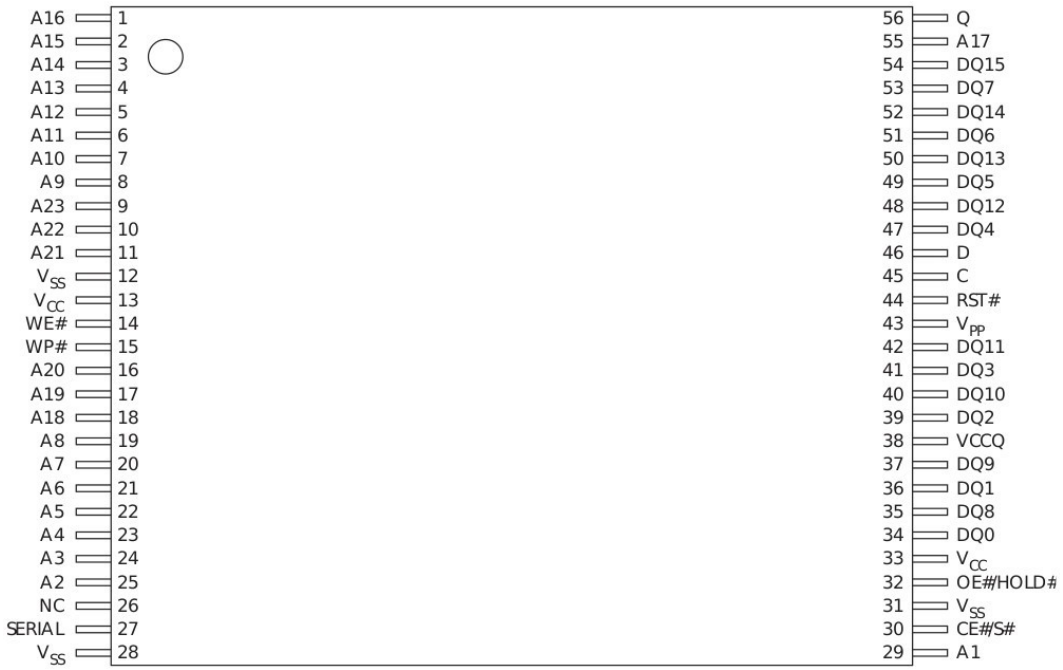
Note: *1 Il Level Voltage Lockout genera un segnale di uscita alto/basso a seconda che il livello della tensione presente sul pin di alimentazione sia sufficiente o meno per eseguire operazioni di scrittura/lettura sul dispositivo di memoria.

Figura 5.6. Dispostivo FeRAM Ramtrom "FM20L08 – 1Mbit Bytewide FRAM Memory Extended Temp." (immagine superiore [14]) e relativa tabella dei Pin del componente di memoria (immagine inferiore). Nota: i dati mostrati in tabella sono stati tratti dal corrispondete datasheet [14].



Nome Pin	Tipo	Funzione
A (17:0)	Input	Pin degli Indirizzi
E	Input	Chip Enable
W	Input	Write Enable
G	Input	Output Enable
LB	Input	Data Lower Byte Enable
UB	Input	Data Upper Byte Enable
DQL (7:0)	Input/Output	Dati I/O (Lower Byte)
DQU (15:8)	Input/Output	Dati I/O (Upper Byte)
V_{DD}	Input	Tensione di Alimentazione
V_{SS}	Input	Massa
NC	---	Pin da Non Collegare

Figura 5.7. Dispostivo MRAM Freescale Semiconductor&Motorola "MR2A16A " (immagine superiore [15]) e "Tabella dei Pin" dello stesso componente di memoria (immagine inferiore). Nota: i dati mostrati in tabella sono stati tratti dal corrispondete datasheet [15].



Nome Pin	Tipo	Funzione
A (MAX:1)	Input	Pin degli Indirizzi
DQ (15:0)	Input/Output	Dati I/O
CE#	Input	Chip Enable
S#	Input (SPI)	SPI Enable
OE#	Input	Output Enable
HOLD#	Input (SPI)	SPI Hold (*4)
RST#	Input	Reset Chip (*1)
WE#	Input	Write Enable
WP#	Input	Write Protect (*2)
C	Input (SPI)	SPI Clock
D	Input (SPI)	SPI Serial Data Input
Q	Output (SPI)	SPI Serial Data Output
SERIAL	Input (SPI)	Ingresso Modalità Seriale/Parallela SPI (*3)
V_{PP}	Input	Tensione Lavoro Operazioni Scrittura/Cancellazione
V_{CC}	Input	Tensione di Alimentazione
V_{CCQ}	Input	Tensione Alimentazione degli Output (*5)
V_{SSQ}	Input	Massa per I/O
V_{SS}	Input	Massa
NC	---	Pin da Non Collegare
DNU	---	Pin da Non Collegare
RFU	---	Pin da Non Collegare

Note: *1 Il Reset Chip consente di attivare/disattivare gli automatismi interni e la possibilità di eseguire operazioni di scrittura sulle celle di memoria;
*2 Il Write Protect attiva/disattiva i meccanismi di "lock-down" del dispositivo;
*3 L'ingresso SERIAL consente di selezionare l'interfaccia Seriale/Parallela della SPI di cui dispone il dispositivo di memoria;
*4 L'ingresso di Hold mette/rimuove l'uscita Q in/dallo stato di alta impedenza;
*5 V_{CCQ} consente a tutti gli output di raggiungere la tensione V_{CCQ} appunto.

Figura 5.8. Dispostivo PCM "P8P Parallel Phase Change Memory (PCM)" Micron Technology (immagine superiore [16]) e "Tabella dei Pin" dello stesso componente di memoria (immagine inferiore). Nota: i dati mostrati in tabella sono stati tratti dal corrispondete datasheet [16].

Riferimenti Bibliografici Introduzione.

[1] http://it.wikipedia.org/wiki/Memoria_(informatica) "Memoria (informatica)", Wikipedia, Wikimedia Foundation;

[2] http://en.wikipedia.org/wiki/Flash_memory "Flash memory", Wikipedia Wikimedia Foundation;

Riferimenti Bibliografici Capitolo1.

[1] it.wikipedia.org/wiki/RAM "RAM", Wikipedia, Wikimedia Foundation;

[2] en.wikipedia.org/wiki/Random-access_memory "Random-access memory", Wikipedia, Wikimedia Foundation;

[3] en.wikipedia.org/wiki/Dynamic_random-access_memory "Dynamic random-access memory", Wikipedia, Wikimedia Foundation;

[4] it.wikipedia.org/wiki/Read_Only_Memory "Read Only Memory", Wikipedia, Wikimedia Foundation;

[5] en.wikipedia.org/wiki/Programmable_read-only_memory "Programmable read-only memory", Wikipedia, Wikimedia Foundation;

[6] it.wikipedia.org/wiki/Antifusibile "Antifusibile", Wikipedia, Wikimedia Foundation;

[7] it.wikipedia.org/wiki/Fusibile "Fusibile", Wikipedia, Wikimedia Foundation;

[8] www-micrel.deis.unibo.it/SMLS/corso/componenti/componenti.htm "Memorie", "Corso di Sistemi a Microprocessore LS Anno Accademico 2007/2008", Bruno Riccò.
Dipartimento di Elettronica, Informatica e Sistemistica, Università di Bologna;

[9] static.usenix.org/events/sec01/full_papers/gutmann/gutmann_html/

"Data Remanence in Semiconductor Devices", Peter Gutmann, IBM T.J.Watson Research Center, Proceedings of the 10th USENIX Security Symposium (Pages 39-54), August 13–17, 2001.
USENIX-The Advanced Computing System Association;
[10] it.wikipedia.org/wiki/EPROM "EPROM", Wikipedia, Wikimedia Foundation;
[11] it.wikipedia.org/wiki/EEPROM "EEPROM", Wikipedia, Wikimedia Foundation;
[12] en.wikipedia.org/wiki/Flash_memory "Flash memory", Wikipedia, Wikimedia Foundation;
[13] it.wikipedia.org/wiki/Memoria_flash "Memoria flash", Wikipedia, Wikimedia Foundation;
[14] "The Zen of Nonvolatile Memories " Pag. 815-820, by Erwin J. Prinz, Freescale Semiconductor, Inc.;
Design Automation Conference 2006, July 24–28, 2006, San Francisco, California, USA.
Copyright 2006 ACM 1-59593-381-6/06/0007;
[15] "Tecnologia e Progettazione di Memorie Non Volatili", A.Pirovano, A. Grossi, R. Bez, G. Servalli.
©2009 Micron Technology, Inc.

Riferimenti Bibliografici Capitolo 2.

[1] "Magnetoresistive Random Access Memory Using Magnetic Tunnel Junctions", S. Tehrani, J.M. Slaugheter, M. Deherrera, B.N. Engel, N.D. Rizzo, J. Salter, M. Durlam, R.W. Dave, J. Janesky, B. Butcher, K. Smith e G. Grynkewich, Invited Paper of Proceedings of the IEEE, Vol. 91, n° 5, MAY 2003.

[2] en.wikipedia.org/wiki/Magnetoresistance "Magnetoresistance", Wikipedia, Wikimedia Foundation;

[3] en.wikipedia.org/wiki/Magnetism "Magnetism", Wikipedia, Wikimedia Foundation;

[4] en.wikipedia.org/wiki/Electron_magnetic_dipole_moment "Electron magnetic dipole moment", Wikipedia, Wikimedia Foundation;

[5] en.wikipedia.org/wiki/Density_of_states "Density of states", Wikipedia, Wikimedia Foundation;

[6] en.wikipedia.org/wiki/Tunnel_magnetoresistance "Tunnel magnetoresistance", Wikipedia, Wikimedia Foundation;

[7] physics.unl.edu/~tsymbal/reference/spin-dependent_tunneling/tunneling_magnetoresistance.shtml "Tunneling Magnetoresistance", Egveny Tsymbal.
Department of Physics and Astronomy – University of Nebraska, Lincoln.

[8] en.wikipedia.org/wiki/Magnetic_anisotropy "Magnetic anisotropy", Wikipedia, Wikimedia Foundation;

[9] http://en.wikipedia.org/wiki/Thermal_fluctuations "Thermal fluctuations", Wikipedia, Wikimedia Foundation;

[10] "The concept and initial studies of a crosstie random access memory (CRAM)", L. J. Schwee, P. E. Hunter, K. A. Restorff, and M. T. Shephard.
J. Appl. Phys., vol. 53, pp. 2762–2764, 1982;

[11] "Fabrication and characterization of a crosstie random access memory" , C. W. Baugh, J. H. Cullom, E. A. Hubbard, M. A. Mentzer, and R. Fedorak.
IEEE Trans. Magn., vol. MAG-18, pp. 1782–1784, Nov. 1982;

[12] "The design of a one megabit nonvolatile M-R memory chip using 1:5 5 µm cells", A. V. Pohm, J. S. T. Huang, J. M. Daughton, D. R. Krahn, and V. Mehra.

IEEE Trans. Magn., vol. 24, pp. 3117–3119, Nov. 1988 ;

[13] “Galvanomagnetic properties of Ag/M (M = Fe, Ni, Co) layered metallic films”, H. Sato, P. A. Schroeder, J. Slaughter, W. P. Pratt, Jr., and W. Abdul- Razzaq.

Superlattices and Microstructures, vol. 4, pp. 45–50, 1988 ;

[14] “Enhanced magnetoresistance of ultrathin (Au/Co)/sub n/ multilayers with perpendicular anisotropy”, E. Velu, C. Dupas, D. Renard, J. P. Renard, and J. Seiden.

Phys. Rev. B, Condens. Matter, vol. 37, p. 668, 1988;

[15] “Giant magnetoresistance of (100)Fe/(001)Cr magnetic superlattices”, M. N. Baibich, J. B. Broto, A. Fert, F. N. Van Day, F. Petroff, P. Etienne, G. Creuset, A. Friedrich, and J. Chazelas.

Phys. Rev. Lett, vol. 61, p. 2472, 1988;

[16] “Enhanced magnetoresistance in layered magnetic structures with antiferromagnetic interlayer exchange”, G. Binasch, P. Grunberg, F. Saurenbach, and W. Zinn.

Phys. Rev. B, Condens. Matter, vol. 39, pp. 4828–4830, 1989;

[17] “Giant magnetic tunneling effect in Fe/Al2O3/Fe junction”, T. Miyazaki and N. Tezuka.

J. Magn. Magn. Mater., vol. 139, p. L231, 1995;

[18] “Large magnetoresistance at room temperature in ferromagnetic thin film tunnel junctions”, J. S. Moodera, L. R. Kinder, T. M. Wong, and R. Meservey.

Phys. Rev. Lett., vol. 74, pp. 3273–3276, 1995;

[19] J. S. Moodera and L. R. Kinider, J. Appl. Phys., vol. 79, p. 4724, 1996.

[20] S. S. P. Parkin, R. E. Fontana, and A. C. Marley, J. Appl. Phys., vol. 81, p. 5521, 1997;

[21] “Large tunneling magnetoresistance enhancement by thermal anneal”,

R. C. Sousa, J. J. Sun, V. Soares, P. P. Freitas, A. Kling, M. F. da Silva, and J. C. Soares.
Appl. Phys. Lett., vol. 73, pp. 3288–3290, 1998;
[22] “Recent developments in magnetic tunnel junction MRAM”, S. Tehrani, B. Engel, J. M. Slaughter, E. Chen, M. DeHerrera, M. Durlam, P. Naji, R. Whig, J. Janesky, and J. Calder.
IEEE Trans. Magn., vol. 36, pp. 2752–2757, Sept. 2000;
[23] “Microstructured magnetic tunnel junctions”, W. J. Gallagher, S. S. P. Parkin, Y. Lu, X. P. Bian, A. Marley, R. A. Altman, S. A. Rishton, K. P. Roche, C. Jahnes, T. M. Shaw, and G. Xiao.
J. Appl. Phys., vol. 81, pp. 3741–3746, 1997;
[24] “Progress and outlook for MRAM technology”, S. Tehrani, J. M. Slaughter, E. Chen, M. Durlam, J. Shi, and M. DeHerrera.
IEEE Trans. Magn., vol. 35, pp. 2814–2819, Sept. 2000;
[25] “Magnetoelectronic memories last and last”, M. Johnson.
IEEE Spectr., vol. 37, pp. 33–40, Feb. 2000.
[26] “A 10 ns read and write non-volatile memory array using a magnetic tunnel junction and FET switch in each cell”, R. Scheuerlein, W. Gallagher, S. Parkin, A. Lee, S. Ray, R. Robertazzi, and W. Reohr.
Solid-State Circuits Conf. Tech. Dig., 2000, pp. 128–129;
[27] “Diode-free magnetic random access memobyr using
spin-dependent tunneling effect”, F. Z. Wang.
Appl. Phys. Lett., vol. 77, no. 13, pp. 2036–2038, 2000;
[28] “Memories of tomorrow”, W. Reohr, H. Honigschmid, R. Robertazzi, D. Gogl, F. Pesavento, S. Lammers, K. Lewis, C. Arndt, Y. Lu, H. Viehmann, R. Scheuerlein, L.-K. Wang, P. Trouilloud, S. Parkin, W. Gallagher, and G. Muller.
IEEE Circuits Devices Mag., vol. 18, no. 5, pp. 17–27, Sept. 2002;

[29] “Inverse tunnel magnetoresistance in Co=SrTiO =La =Sr =MnO : New ideas on spin-polarized tunneling”, J. M. de Teresa, A. Barthelemy, A. Fert, J. P. Contour, R. Lyonnet, F. Montaigne, P. Seneor, and A. Vaures.
Phys. Rev. Lett., vol. 82, pp. 4288–4291, 1999;
[30] “Fundamentals of MRAM technology” , J. M. Slaughter, R. W. Dave, M. DeHerrera, M. Durlam, B. N. Engel, N. D. Rizzo, and S. Tehrani.
J. Superconduct., vol. 15, no. 1, pp. 19–25, 2002;
[31] “Low-resistance spin-dependent tunnel junctions with ZrAlO barriers”, J.Jianguo Wang, P. P. Freitas, and E. Snoeck.
Appl. Phys. Lett., vol. 79, no. 27, pp. 4553–4555, Dec. 2001;
[32] “A mechanism of magnetic hysteresis in heterogeneous alloys”, E. C. Stoner and E. P. Wohlfarth.
Philos. Trans. R. Soc. London A, Math. Phys. Sci., vol. A-240, pp. 599–642, 1948;
[33] “End domain states and magnetization reversal in submicron magnetic structures”, J. Shi, T. Zhu, M. Durlam, E. Chen, S. Tehrani, Y. F. Zheng, and J.-G. Zhu.
IEEE Trans. Magn., vol. 34, p. 997, July 1998;
[34] “Magnetization vortices and anomalous switching in patterned NiFeCo submicron arrays”, J. Shi, S. Tehrani, T. Zhu, Y. F. Zheng, and J. G. Zhu.
Appl. Phy. Lett., vol. 74, pp. 2525–2527, 1999;
[35] “Geometry dependence of magnetization vortices in patterned submicron NiFe elements”, J. Shi, S. Tehrani, and M. Scheinfein.
Appl. Phys. Lett., vol. 76, pp. 2588–2590, 2000;
[36] “Edge pinned states in patterned submicron ultra-thin film magnetic structures,”J. Shi and S. Tehrani.
Appl. Phys. Lett., vol. 77, pp. 1692–1694, 2000;
[37] “Thermally activated magnetization reversal in submicron magnetic

tunnel junctions for magnetoresistive random access memory", N. D. Rizzo, M. DeHerrera, J. Janesky, B. Engel, J. Slaughter, and S. Tehrani.
Appl. Phys. Lett., vol. 80, pp. 2335–2337, 2002;
[38] "Magnetostatic interactions between submicrometer patterned magnetic elements", J. Janesky, N. D. Rizzo, L. Savtchenko, B. Engel, J. M. Slaughter, and S. Tehrani.
IEEE Trans. Magn., vol. 37, pp. 2052–2054, July 2001;
[39] "High-speed characterization of submicrometer giant magnetoresistive devices" J. Appl. Phys., vol. 85, pp. 4773–4775, 1999;
[40] "Magnetization reversal in micron-sized magnetic thin film", R. H. Koch, J. G. Deak, D. W. Abraham, P. L. Trouilloud, R. A. Altman, Y. Lu, W. J. Gallagher, R. E. Scheuerlein, K. P. Roche, and S. S. P. Parkin,
Phys. Rev. Lett., vol. 81, p. 4512, 1998;
[41] "A 256 kb 3.0 V 1T1MTJ nonvolatile magnetoresistive RAM", P. K. Naji, M. Durlam, S. Tehrani, J. Calder, and M. F. DeHerrera.
IEEE ISSCC Dig. Tech. Papers, vol. 438, Feb. 2001, pp. 122–123;

Riferimenti Bibliografici Capitolo 3.

[1] "A Survey of Circuit Innovations in Ferroelectric Random-Access Memories", Ali Sheikholeslami and P. Glenn Gulak.
Proceedings of IEEE, Vol. 88, n°5, pp. 667-689, May 2000.
Publisher Item Identifier S 0018-9219(00)04568-0. © 2000 IEEE ;
[2] "Trends in Electronic Reliability – Effect of Terrestrial Cosmic Rays", J. F. Ziegler – SRIM&TRIM www.srim.org;
[3] "Non Volatile Random Access Memory Technologies (MRAM, FeRAM, PRAM)", Muhammad Muneeb, Imran Akram e Aftab Nazir, www.imit.kth.se/info/SSD/KMF/2B1750/2B1750_06_RAMs.pdf.

KTH Royal Institute of Technology, Stoccolma;

[4] it.wikipedia.org/wiki/Perovskiti "Perovskiti", Wikipedia, Wikimedia Foundation;

[5] www.scribd.com/doc/62391239/Nvram-Presentation "NON-VOLATILE RANDOM ACCESS MEMORY (NVRAM)", Kaustav Roy, pp. 1-22, August 16, 2011;

[6] "Nonvolatile multilevel memories for digital applications", B. Ricco, G. Torelli, M. Lanzoni, A. Manstretta, H. Maes, D. Montanari, and A. Modelli.
Proc. IEEE, vol. 86, pp. 2399–2421, Dec. 1998;

[7] "Flash memory cells – An overview", P. Pavan, R. Bez, P. Olivi, and E. Zanoni.
Proc. IEEE, vol. 85, pp. 1248–1271, Aug. 1997;

[8] "Ferroelectric nonvolatile memories for embedded applications", R. E. Jones Jr.
Proc. IEEE Custom Integrated Circuits Conf., 1998, pp. 431–438;

[9] "An embedded FeRAM macro cell for a smart card microcontroller", T. Miwa, J. Yamada, Y. Okamoto, H. Koike, H. Toyoshima, H. Hada, Y. Hayashi, H. Okizaki, Y. Miyasaka, T. Kunio, H. Miyamoto, H. Gomi, and H. Kitajima.
Proc. IEEE Custom Integrated Circuits Conf., 1998, pp. 439–442;

[10] "Endurance properties of ferroelectric PZT thin films", R. Moazzami, C. Hu, and W. H. Shepherd.
Tech. Dig. IEEE Int. Electron Devices Meeting, Dec. 1990, pp. 181–184;

[11] "Impact of polarization relaxation on ferroelectric memory performance", R. Moazzami, N. Abt, Y. Nissan-Cohen, W. H. Shepherd, M. P. Brassington, and C. Hu.
Tech. Dig. Symp. VLSI Technology, May 1991, pp. 61–62;

[12] “Imprint of ferroelectric PLZT thin-film capacitors with lanthanum strontium cobalt oxide electrodes”, J. M. Benedetto, M. L. Roush, I. K. Lloyd, and R. Ramesh.
Proc. 9th Int. Symp. Appl. of Ferroelectrics, 1994, pp. 66–69;
[13] “A survey of magnetic and other solid-state devices for the manipulation of information”, J. A. Rajchman.
IRE Trans. Circuit Theory, pp. 210–225, Sept. 1957;
[14] Polar Dielectrics and Their Applications, J. C. Burfoot and G. W. Taylor.
Berkeley, CA: Univ. of California Press, 1979;
[15] “A 42.5 mm 1 Mb nonvolatile ferroelectric memory utilizing advanced architecture for enhanced reliability”, W. Kraus, L. Lehman, D. Wilson, T. Yamazaki, C. Ohno, E. Nagai, H. Yamazaki, and H. Suzuki.
Symp. VLSI Circuits Dig. Tech. Papers, 1998, pp. 242–245;
[16] “A 256 kb nonvolatile ferroelectric memory at 3 V and 100 ns”,T. Sumi, N. Moriwaki, G. Nakane, T. Nakakuma, Y. Judai, Y. Uemoto, Y. Nagano, S. Hayashi, M. Azuma, E. Fujii, S. Katsu, T. Otsuki, L. McMillan, C. P. de Araujo, and G. Kano.
ISSCC Dig. Tech. Papers, 1994, pp. 268–269;
[17] “Folded bit line ferroelectric memory device”, T. A. Lowrey and W. L. Kinney.
U.S. Patent 5 541 872, July 30, 1996;
[18] “Signal magnitudes in high density ferroelectric memories”, W. Kinney.
Integrat. Ferroelect., vol. 4, pp. 131–144, 1994;
[19] “A 16 kb ferroelectric nonvolatile memory with a bit parallel architecture”, R. Womack and D. Tolsch.
ISSCC Dig. Tech. Papers, 1989, pp. 242–243;
[20] “Ferroelectric nonvolatile random access memory having drive line

segments", R. E. Jones Jr..
U.S. Patent 5 373 463, Dec. 13, 1994;
[21] "Multi-phase driven split word line ferroelectric memory without PL", H. B. Kang, D. M. Kim, K. Y. Oh, J. S. Roh, J. J. Kim, J. H. Ahn, H. G. Lee, D. C. Kim, W. Jo, H. M. Lee, S. M. Cho, H. J. Nam, J. W. Lee, and C. S. Kim,
ISSCC Dig. Tech. Papers, 1999, pp. 108–109 ;
[22] "A 60 ns 1 Mb nonvolatile ferroelectric memory with non-
driven cell plate line write/read scheme", H. Koike, T. Otsuki, T. Kimura, M. Fukuma, Y. Hayashi, Y. Maejima, K. Amanuma, M. Tanabe, T. Matsuki, S. Saito, T. Takeuchi, S. Kobayashi, T. Kunio, T. Hase, Y. Miyasaka, N. Shohota, and M. Takada.
ISSCC Dig. Tech. Papers, 1996, pp. 368–369;
[23] "Memory cell with volatile and non-volatile portions having ferroelectric capacitors", K. Dimmler and S. Eaton.
U.S. Patent 4 809 225, Feb. 28, 1989;
[24] "Novel gain cell with ferroelectric coplanar capacitor for high-density nonvolatile random-access memory", M. Aoki, H. Takauchi, and H. Tamura.
Tech. Dig. IEEE Int. Electron Devices Meeting, 1997, pp. 942–944;
[25] "High-density chain ferroelectric random access memory (chain FRAM)", D. Takashima and I. Kunishima.
IEEE J. Solid-State Circuits, vol. 33, pp. 787–792, May 1998;
[26] "A sub-40 ns random-access chain FRAM architecture with a 7 ns cell-plate-line drive", D. Takashima, S. Shuto, I. Kunishima, H. Takenaka, Y. Oowaki, and S. Tanaka.
ISSCC Dig. Tech. Papers, 1999, pp. 102–103;

Riferimenti Bibliografici Capitolo 4.

[1] “Architecting Phase Change Memory as a Scalable DRAM Alternative” (Copyright 2009 ACM 978-1-60558-526-0/09/06), Benjamin C. Lee, Engin Ipek, Onur Mutlu e Doug Burger.
Internationl Symposium Computer Architecture, 20-24 Giugno 2009.
[2] “Highly manufacturable high density phase change memory of 64Mb and beyond”, S. Ahn et al.
International Electron Devices Meeting, 2004.
[3] Figura tratta da “Phase-Change Memory”, H.-S. Philip Wong, Simone Raoux, SangBum Kim, Jiale Liang, John P. Reifenberg, Bipin Rajendran, Mehdi Asheghi e Kenneth E. Goodson., collaborazione di Stanford University, IBM T.J. Watson Research Center ed Intel Corporation.
Proceedings of the IEEE, Vol. 98, No. 12, Dicembre 2010.
[4] it.wikipedia.org/wiki/Elementi_del_gruppo_16 “Elementi del gruppo 16”, Wikipedia, Wikimedia Foundation;
[5] “Process integration, devices & structures”, International Technology Roadmap for Semiconductors, 2007.
[6] “An 8Mb demonstrator for high-density 1.8V phase-change memories”, F. Bedeschi et al.
Symposium on VLSI Circuits, 2004.
[7] “A multi-level-cell bipolar-selected phase-change memory”, F. Bedeschi et al.
International Solid-State Circuits Conference, 2008.
[8] “Ultra-thin phase-change bridge memory device using GeSb”, Y. Chen et al.
International Electron Devices Meeting, 2006.
[9] “High-performance and low-voltage sense-amplifier techniques for sub-

90nm sram”, M. Sinha et al.
International Systems-on-Chip Conference, 2003.
[10] “CMOS VLSI Design”, N. Weste and D. Harris.
Pearson Education”, 3rd edition, 2005.
[11] “A novel cell technology using N-doped GeSbTe films for phase change RAM”, H. Horii et al.
Symposium on VLSI Technology, 2003.
[12] “A 0.1um 1.8V 256Mb 66MHz synchronous burst PRAM”, S. Kang et al.
International Solid-State Circuits Conference, 2006.
[13] “FlashCache: A NAND flash memory file cache for low power web servers”, T. Kgil and T. Mudge.
International Conference on Compilers, Architecture, and Synthesis for Embedded Systems, October 2006.
[14] “Current status of the phase change memory and its future”, S. Lai.
International Electron Devices Meeting, 2003.
[15] “A 90nm 1.8V 512Mb diode-switch PRAM with 266 MB/s read throughput”, K.-J. Lee et al.
Journal of Solid-State Circuits, 43(1), January 2008.
[16] “512Mb DDR2 SDRAM component data sheet: MT47H128M4B6-25”. Micron.
www.micron.com, March 2006.
[17] “Technical note TN-47-04: Calculating memory system power for DDR2”. Micron.
www.micron.com, June 2006.
[18] “Optimizing NUCA organizations and wiring alternatives for large caches with CACTI 6.0”, N. Muralimanohar et al.
International Symposium on Microarchitecture, December 2007.

[20] “Enhanced write performance of a 64mb phase-change random access memory”, H. Oh et al.
International Solid-State Circuits Conference, 2005.
[21] “A 90nm phase change memory technology for stand-alone non-volatile memory applications”, F. Pellizzer et al.
Symposium on VLSI Circuits, 2006.
[22] “Scaling analysis of phase-change memory technology”, A. Pirovano et al.
International Electron Devices Meeting, 2003.

Riferimenti Bibliografici Capitolo5.

[1] “Process Integration, Devices, and Structures ”, International Technlogy Roadmap for Semiconductors, 2007 Edition.
[2] “System Drivers”, tabella “Embedded Memory Requirements”, International Technlogy Roadmap for Semiconductors, 2007 Edition.
[3] “System Drivers”, sezione “Embedded Applications” della tabella “Embedded Memory Requirements”, International Technlogy Roadmap for Semiconductors, 2007 Edition.
[4] “Full-Bit Functional, High-Density 8 Mbit One Transistor-One Capacitor Ferroelectric Random Access Memory Embedded Within A Low-Power 130 nm Logic Process”, K. R. Udayakumar et al.
Jap. J. Appl. Phys. 46 (2007) 2180-2183.
[5] “Nanoelectronics and Information Technology”, Ed. Rainer Waser. Wiley-VCH, 2003, 568-569.
[6] “Current Status And Challenges Of Ferroelectric Memory Devices”, H. Kohlstedt et al.
Microelectronic Eng. 80 (2005) 296- 304.

[7] "MRAM Cell Technology For Over 500-MHz SOC", N. Sakimura et. al. IEEE J. Solid-State Circ. 42 (2007) 830-838.

[8] "Ballistic bit addressing in a magnetic memory cell array", H. W. Schumacher.

Appl. Phys. Lett. v. 87 , no. 4 (2005) 42504.

[9] "A 0.18-mm 3.0-V 64-Mb Nonvolatile Phase-Transition Random Access Memory (PRAM)", W. Y. Cho, B-H Cho, B-G. Choi, H-R Oh, S. Kang, K-S. Kim, K-H. Kim, D-E. Kim, C-K. Kwak, H-G. Byun, Y. Hwang, S. J. Ahn,

G-H. Koh, G. Jeong. H. Jeong, and K. Kim.

IEEE J. Solid-State Circuits v. 40, no. 1 (2005) 291-300.

[10] "Assessment of the Potential & Maturity of Selected Emerging Research Memory Technologies ", Jim Hutchby, Mike Garner .

Workshop & ERD/ERM Working Group Meeting (April 6-7, 2010).

[11] en.wikipedia.org/wiki/Phase-change_memory "Phase-change memory", Wikipedia, Wikimedia Foundation.

[12] en.wikipedia.org/wiki/MRAM "Magnetoresistive random-access memory", Wikipedia, Wikimedia Foundation.

[13] en.wikipedia.org/wiki/Ferroelectric_RAM "Ferroelectric RAM", Wikipedia, Wikimedia Foundation.

[14] "FM20L08 1Mbit Bytewide FRAM Memory – Extended Temp", Rev. 1.4. October 2005, Ramtron International Corporation , 1850 Ramtron Drive, Colorado Springs, CO 80921.

(800) 545-FRAM, (719) 481-7000 .

[15] "Freescale Semiconductor Data Sheet 256Kx16-Bit 3.3-V Asynchronous Magnetoresistive RAM", Document Number: MR2A16A, Rev. 6, 11/2007.

[16] “P8P Parallel Phase Change Memory (PCM) ”, PDF: 09005aef8447d46d/Source: 09005aef845b5c96
parallel_pcm_1.fm - Rev. J 11/11 EN .

Ringraziamenti

- Si ringrazia infinitamente il Professor Dott. Ing. Aldo Romani per il preziosissimo contributo fornitomi nella stesura di questo elaborato finale, e per la grande disponibilità e pazienza mostratami costantemente lungo tutto il periodo di tempo dedicato all'elaborazione di questo testo.

- Si ringraziano i miei genitori ed i miei nonni, che non hanno mai smesso di supportarmi economicamente, fornirmi il loro indispensabile sostegno psicologico nei momenti meno esaltanti, e che hanno sempre assecondato tutte le mie scelte nell'ambito di questo percorso formativo. Verso di essi sono quindi indescrivibilmente riconoscente perché a loro devo semplicemente tutto.

- Si ringrazia il Dott. Ing. Aerospaziale Tomas Michael Tesfu, che nel corso di questi anni trascorsi a Cesena si è dimostrato essere un fratello, prima ancora che un fidatissimo amico, prima ancora che una simpatica canaglia.
Lo ringrazio per tutti i fondamentali consigli forniti in questo lungo periodo di tempo, per avermi aiutato a risolvere la maggior parte dei problemi quotidiani che caratterizzano la vita uno studente universitario fuori sede, e per avermi fatto sentire a casa sin dal primo giorno della mia permanenza in Romagna.

- Si ringrazia il Comandante dell'Aviazione Civile Italiana Nichi Ranaldo, che con il suo elevato profilo professionale, ha rappresentato per me un modello cui ispirarsi per quanto riguarda l'ampliamento delle proprie conoscenze.

- Si ringrazia il Dott. in Criminologia Applicata Fabio De Vitis, per la grande pazienza e l'estrema generosità dimostratami nell'ospitarmi a Forlì senza contropartita alcuna in questi miei ultimi mesi di studio.

- Si ringrazia tutti i parenti ed amici, ed in particolar modo la sempre cordiale e paziente Annamaria Mele, per il loro sincero e costante sostegno che mi hanno offerto nel corso di cinque questi anni.