UNIVERSITÀ DEGLI STUDI DI PADOVA

DIPARTIMENTO DI INGEGNERIA

DIPARTIMENTO DI INGEGNERIA DELL'INFORMAZIONE

CORSO DI LAUREA IN INGEGNERIA INFORMATICA

L'EVOLUZIONE DELLE MEMORIE NON VOLATILI

Relatore:
Prof. Zanoni Enrico

Laureando:
Munerotto Giacomo

ANNO ACCADEMICO 2025-2026

Data di laurea 12/03/2026

Abstract

Negli ultimi decenni, la continua evoluzione dei sistemi digitali ha portato a una crescente richiesta di memorie più veloci, efficienti e affidabili. Sebbene le memorie volatili come la DRAM abbiano dominato per anni molte applicazioni, le memorie non volatili (NVM) stanno assumendo un ruolo sempre più centrale, non solo nello storage ma anche nell'elaborazione dati. Accanto alle tecnologie consolidate come EEPROM e Flash, si stanno affermando nuove soluzioni emergenti, tra cui: RRAM, MRAM, PCM e FRAM, capaci di offrire prestazioni superiori in termini di velocità, consumo energetico e durata.

Questa tesi si propone di analizzare l'evoluzione delle memorie non volatili, con particolare attenzione alle tecnologie emergenti. Dopo aver introdotto le definizioni, le classificazioni e le caratteristiche delle memorie in generale, viene presentato un confronto tecnico tra memorie volatili e non volatili. Si analizzano quindi le principali tecnologie emergenti, il loro stato dell'arte e le loro possibili applicazioni nei settori mobili, embedded, datacenter e intelligenza artificiale. Infine, viene proposta un'analisi comparativa tra le diverse NVM emergenti, evidenziandone le potenzialità, le sfide ancora aperte e le prospettive future.

Indice

Introduzione 1

Le memorie: definizioni e classificazioni 2

1.1 Definizione di memoria in ambito informatico 2

1.2 Classificazione generale delle memorie 4

1.3 Caratteristiche tecniche e prestazionali delle memorie 5

Memorie volatili vs Memorie non volatili 9

2.1 Memorie volatili (RAM, DRAM, SRAM): caratteristiche e utilizzi 9

2.2 Memorie non volatili (ROM, EEPROM, FLASH, PCM): caratteristiche e utilizzi 12

2.3 Confronto tecnico e funzionale tra VM e NVM 16

Evoluzione delle memorie non volatili 17

3.1 RRAM and MRAM 17

3.1.1 RRAM 17

3.1.2 MRAM 19

3.2 Phase Change Memory (PCM) e OxRAM (Oxide-based Resistive RAM) 21

3.2.1 Phase Change Memory 21

3.2.2 OxRRAM 23

3.3 PCM e OxRRAM come sinapsi nelle reti neurali spiking 25

3.3.1 OxRAM come sinapsi artificiale 25

3.3.2 PCM come sinapsi artificiale bidirezionale 28

3.3.4 Conclusioni 31

3.4 STT-RAM come una memoria universale 32

Applicazioni e integrazione delle NVM emergenti 37

4.1 Utilizzo nei dispositivi embedded e Iot 37

4.1.1 STT-MRAM Deep Learning Model for IoT Applications 38

4.2 Integrazione nelle architetture di sistema (memorie artificiali, neuromorphic computing) 41

4.3 Impatto su datacenter, HPC e applicazioni AI 42
4.3.1 Architettura di acceleratore AI ottimizzato con STT-MRAM 43
Analisi comparativa e prospettive future 45
5.1 Confronto tra le principali tecnologie NVM emergenti 45
5.1.1 Analisi dei parametri prestazionali 45
5.1.2 Verso una specializzazione delle tecnologie 46
5.2 Sfide tecnologiche e limiti attuali 47
5.3 Prospettive evolutive e roadmap tecnologiche 47
Conclusioni 49
6.1 Sintesi dei principali risultati 49
6.2 Considerazioni sull'impatto della nuova tecnologia 50
6.3 Spunti per future richerche 51
Bibliografia 52

Introduzione

La memoria è uno degli elementi fondamentali di ogni sistema informatico. Essa consente la memorizzazione temporanea o permanente delle informazioni necessarie al funzionamento dei dispositivi digitali, dai semplici microcontrollori agli elaboratori ad alte prestazioni. La distinzione tra memorie volatili e non volatili rappresenta una delle principali classificazioni, con implicazioni significative dal punto di vista strutturale, prestazionale ed energetico.

Con l'aumento della complessità dei sistemi e la diffusione di tecnologie come l'Internet of Things (IoT), il calcolo ad alte prestazioni (HPC) e l'intelligenza artificiale, si è resa necessaria l'introduzione di nuove soluzioni di memoria capaci di superare i limiti delle tecnologie tradizionali. In questo contesto, le memorie non volatili emergenti come la Resistive RAM (RRAM), la Magnetoresistive RAM (MRAM), la Phase Change Memory (PCM) e la Ferroelectric RAM (FRAM), rappresentano una risposta concreta e innovativa.

Lo scopo di questa tesi è duplice: da un lato fornire una panoramica chiara e strutturata dell'evoluzione delle memorie non volatili, evidenziandone i principi di funzionamento e lo stato dell'arte; dall'altro analizzarne le applicazioni più promettenti, i limiti tecnologici attuali e le possibili traiettorie evolutive.

L'analisi si conclude con una riflessione sull'impatto che tali tecnologie potrebbero avere sui futuri sistemi computazionali e con alcuni spunti di approfondimento per la ricerca futura.

Capitolo 1

Le memorie: definizioni e classificazioni

1.1 Definizione di memoria in ambito informatico

Nel contesto informatico, la memoria è un componente fondamentale che consente a un sistema di elaborazione di immagazzinare, recuperare e modificare dati e istruzioni. Essa rappresenta una delle tre unità funzionali principali di un computer, insieme al processore (CPU) e ai dispositivi di input/output. Le memorie svolgono un ruolo cruciale nel determinare le prestazioni complessive del sistema, poiché il tempo impiegato per accedere ai dati può influenzare significativamente la velocità di esecuzione dei programmi. Una memoria, a livello base, è costituita da celle di memoria, ciascuna delle quali può contenere un dato binario (0 o 1). Ogni cella è identificata da un indirizzo univoco, che consente alla CPU di localizzare rapidamente i dati desiderati. A seconda della loro funzione, le memorie possono contenere dati temporanei, come durante l'esecuzione di un programma, oppure permanenti, come il contenuto di un sistema operativo. Possiamo distinguere la memoria in primaria e secondaria vedi Fig. 1.1. La memoria primaria (come la RAM) è direttamente accessibile dalla CPU e serve a mantenere i dati in uso, la memoria secondaria (come dischi rigidi o SSD) conserva i dati in modo permanente ma con tempi di accesso più lunghi.

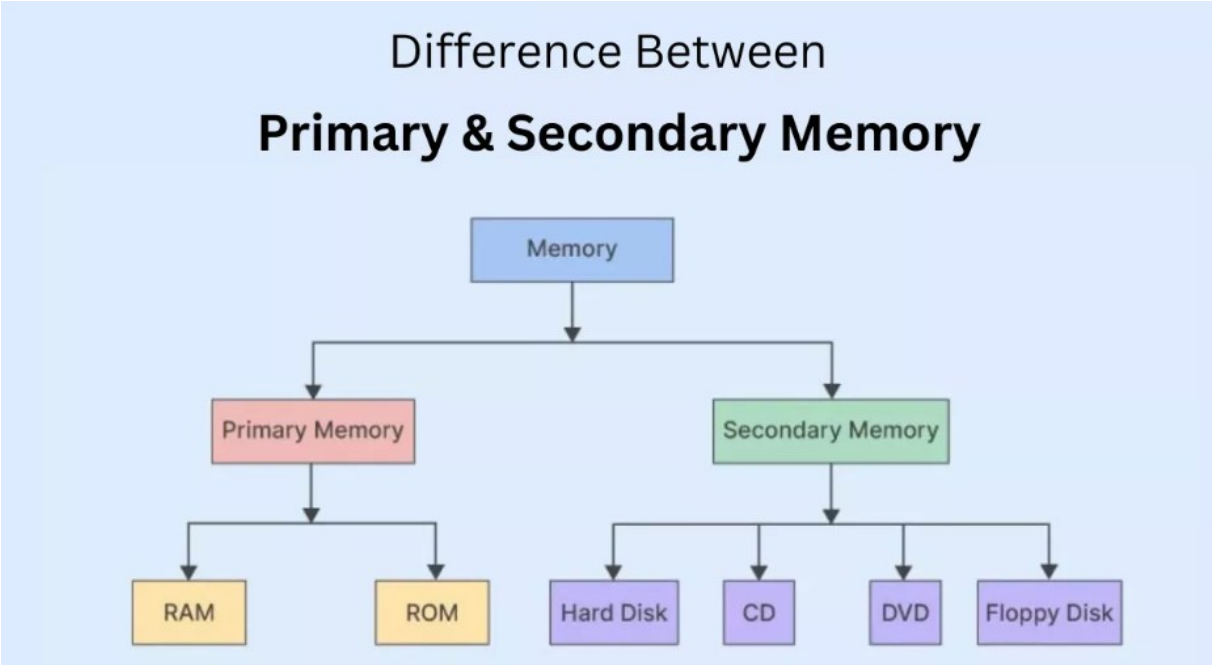


Figura 1.1: A sinistra tipologie di memorie primarie e a destra le memorie secondarie

Differenziamo le memorie anche in volatili e non volatili: le prime perdono il loro contenuto allo spegnimento del sistema, mentre le seconde lo mantengono anche in assenza di alimentazione. A seconda delle esigenze applicative, si preferisce un tipo all'altro in base a parametri quali velocità, persistenza, capacità e costo. In sintesi, la memoria in informatica è ciò che consente ad un sistema di "ricordare" e manipolare le informazioni necessarie al suo funzionamento. Comprendere le sue caratteristiche e il suo ruolo è essenziale per analizzare le evoluzioni tecnologiche che hanno portato allo sviluppo delle moderne memorie non volatili.

1.2 Classificazione generale delle memorie

Le memorie in un sistema informatico si classificano in base a diversi criteri funzionali, architetturali e tecnologici. Questa classificazione è fondamentale per comprendere il ruolo specifico che ciascun tipo di memoria svolge all'interno dell'architettura del calcolatore.

Le **memorie primarie** sono quelle direttamente accessibili dalla CPU e sono utilizzate per conservare istruzioni e dati temporanei durante l'esecuzione dei programmi. Appartengono a questa categoria:

-RAM (Random Access Memory): memoria volatile ad accesso casuale, utilizzata per caricare i programmi e i dati in uso. Può essere di tipo DRAM (più lenta ed economica della SRAM) o SRAM (più veloce e costosa).

-ROM (Read Only Memory): memoria non volatile contenente dati permanenti, solitamente utilizzata per firmware. Ne esistono varianti riscrivibili come EPROM, EEPROM e flash ROM.

-Cache: memoria intermedia ad alta velocità situata tra la CPU e la RAM. Organizzata in livelli (L1, L2, L3), ha lo scopo di ridurre la latenza di accesso ai dati più frequentemente utilizzati. Sono fondamentali per colmare il divario di velocità tra CPU e memoria principale.

Le **memorie secondarie** non sono direttamente accessibili dalla CPU, ma servono per l'archiviazione permanente o semi-permanente di grandi quantità di dati. Hanno capacità elevate, ma velocità inferiori rispetto alle memorie primarie:

-Hard Disk Drive (HDD): memoria magnetica meccanica, economica e con elevata capacità, ma con tempi di accesso più lenti.

-Solid State Drive (SSD): memoria a stato solido basata su tecnologia Flash, più veloce degli HDD e priva di parti mobili, ma con un costo per bit superiore.

-Memorie Flash: usate solitamente in chiavette USB, schede SD. Offrono un buon compromesso tra capacità, velocità e persistenza.

Le **memorie di massa** sono utilizzate per archiviare grandi volumi di dati non frequentemente accessibili, mentre le **memorie di backup** sono pensate per il salvataggio sicuro dei dati nel tempo. La classificazione delle memorie si integra nel concetto di gerarchia della memoria, in cui ogni livello offre un compromesso tra velocità, capacità e costo. I livelli più vicini alla CPU sono più rapidi ma limitati in capacità; quelli più lontani sono più lenti ma più economici e capienti.

1.3 Caratteristiche tecniche e prestazionali delle memorie

In questo sottocapitolo si analizzeranno le principali caratteristiche tecniche e prestazionali delle memorie utilizzate nei sistemi informatici. Verranno esaminati parametri fondamentali quali latenza, velocità di trasferimento, persistenza, costo per bit e capacità, elementi chiave per comprendere le differenze tra le varie tecnologie e il loro impatto sull'efficienza complessiva del sistema.

Bisogna innanzitutto dare una delucidazione sulla gerarchia della memoria (vedi Fig. 1.2) brevemente introdotta alla fine del precedente sottocapitolo [1.2]. Essa è un'organizzazione a livelli che consente ai sistemi informatici di bilanciare velocità, costo e capacità delle memorie. La base di una gerarchia di memoria è il principio di **località**, che afferma che la maggior parte dei programmi non accede in modo uniforme ai dati. Esistono due tipi di località:

-Località Temporale: Gli elementi a cui si è acceduto di recente verranno probabilmente utilizzati nuovamente nell'imminente futuro.

-Località Spaziale: Gli elementi i cui indirizzi sono vicini tendono ad essere referenziati insieme nel tempo.

Questo principio, insieme all'idea che l'hardware più piccolo può essere reso più veloce, ha portato alla creazione di gerarchie basate su memorie di diverse velocità e dimensioni. Con hardware più piccolo si fa riferimento allo *scaling* ossia la riduzione delle dimensioni dei componenti elettronici dei circuiti, in particolare dei transistor.

Quando un transistor diventa più piccolo:

1) La **velocità** aumenta: la corrente elettrica deve percorrere distanze più brevi; quindi, i segnali viaggiano più velocemente all'interno del chip.

2) Il **consumo energetico** diminuisce: con dimensioni minori, la capacità elettrica tra le varie parti diminuisce, riducendo l'energia necessaria per commutare i transistor.

3) Si possono inserire **più transistor** nello stesso spazio: ciò permette di avere circuiti più complessi o più unità di memoria vicine al processore senza aumentare le dimensioni fisiche del chip.

Una gerarchia di memoria è organizzata in diversi livelli. Ogni livello è più piccolo, più veloce e più costoso per byte rispetto al livello inferiore successivo, che è più lontano dal

processore. L'obiettivo è ottenere un sistema di memoria con un costo per byte simile a quello del livello più economico e una velocità quasi pari a quella del livello più veloce.

Gestione dei Miss e delle Penalties: Quando una word non viene trovata nella cache si ha un *cache miss*, deve essere perciò recuperata da un livello inferiore della gerarchia (potrebbe essere un'altra cache o la memoria principale) e posizionata nella cache prima di continuare.

Per ridurre le *penalties* si adotta la strategia di aggiungere cache multilivello (L1, L2, L3) utilizzate per catturare accessi che altrimenti andrebbero alla memoria principale. Le cache multilivello sono inoltre più efficienti dal punto di vista energetico rispetto ad una singola cache aggregata. [2]

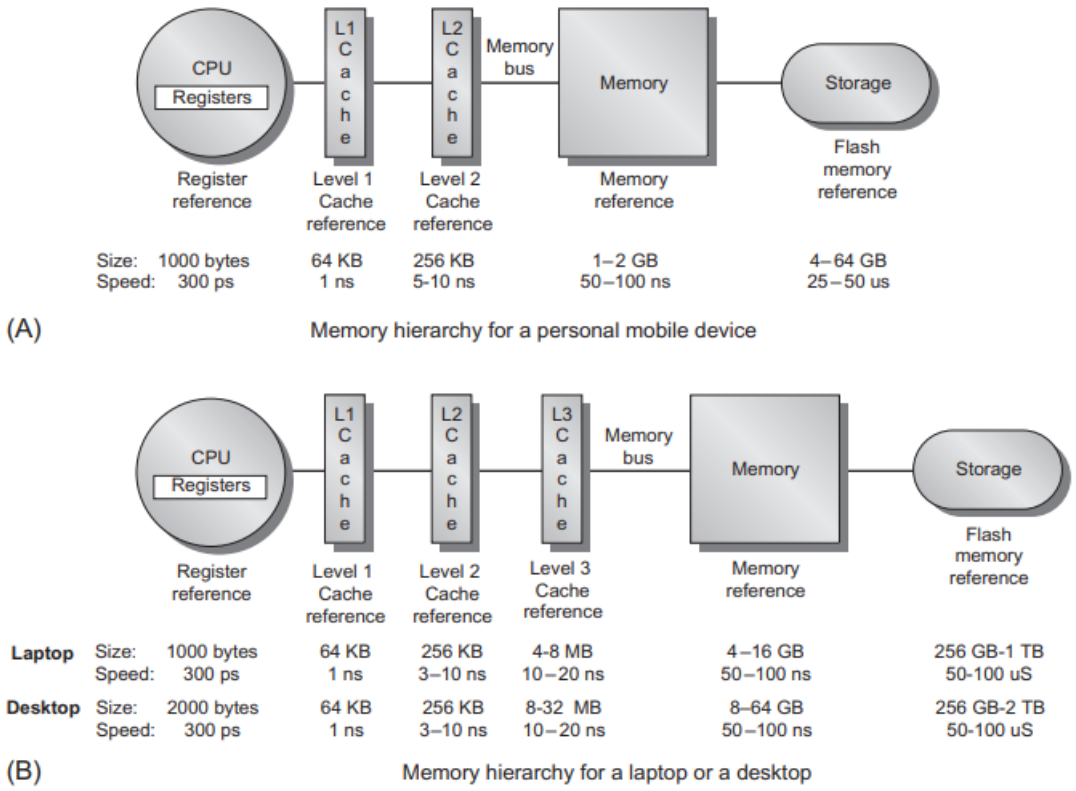


Figura 1.2: A) I livelli in una tipica memoria gerarchica in un telefono o tablet, B) Laptop

La **latenza** indica il tempo necessario per accedere ad un dato elemento di memoria a partire dal momento in cui la richiesta viene fatta. In una gerarchia di memoria, la latenza varia notevolmente tra i diversi livelli: la cache di primo livello (L1) ha una latenza molto bassa (pochi cicli di clock), mentre la memoria principale (RAM) ha una latenza più elevata, spesso di un ordine di grandezza superiore. Il tempo totale per accedere alla memoria in un sistema con gerarchia è dato dalla formula dell'**AMAT** (Average Memory Access Time), che combina

la latenza della cache e quella della memoria principale, tenendo conto della percentuale di accessi che risultano in un *cache hit* o un *cache miss*:

AMAT = Hit time + Miss rate x Miss penalty [1]

-Hit time: tempo per accedere alla cache se il dato è presente (cache hit)

-Miss rate: probabilità che il dato richiesto non sia in cache (cache miss)

-Miss penalty: tempo aggiuntivo necessario per recuperare il dato dalla memoria principale in caso di miss.

La **velocità** (throughput) in ambito informatico, e in particolare nel contesto delle memorie indica la quantità di dati che un sistema può trasferire in un determinato intervallo di tempo. Il *throughput* è influenzato dalla località che abbiamo analizzato in precedenza, e dalla capacità dei diversi livelli della gerarchia di memoria. Ad esempio, nella *Memory Mountain*, il *throughput* varia di oltre un ordine di grandezza a seconda che i dati siano stati usati dalla cache L1 o dalla memoria principale.

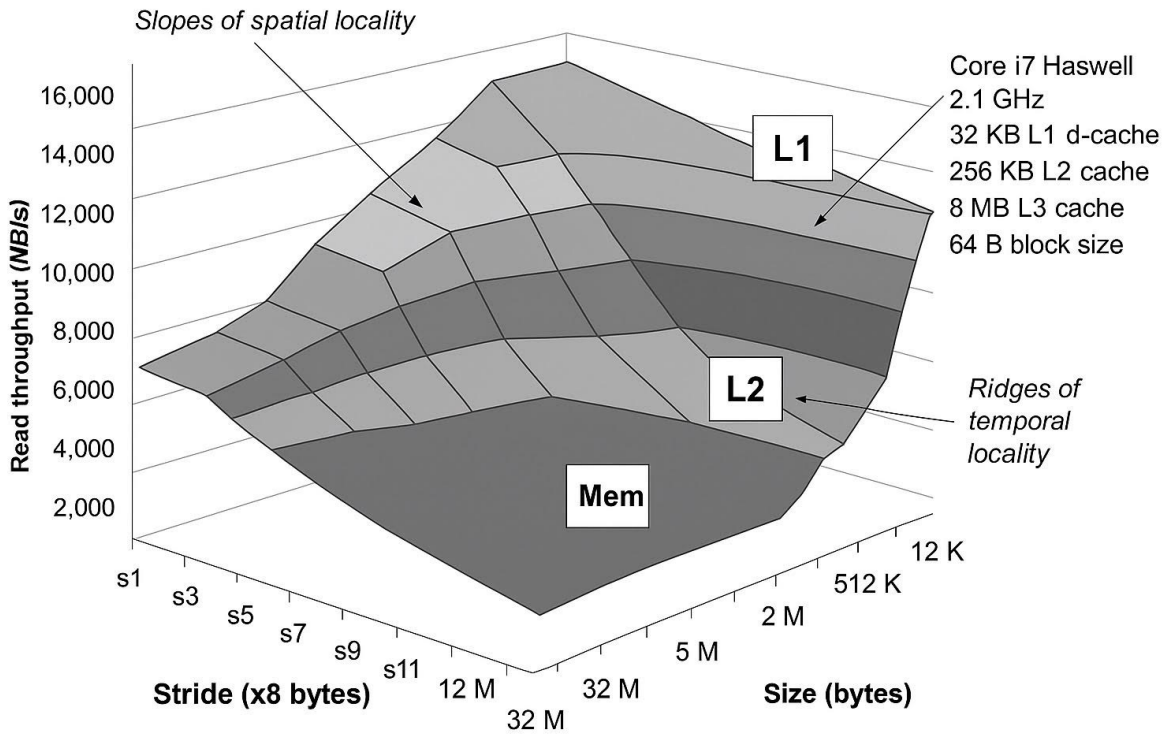


Figura 1.3: La Memory Mountain che mostra il read throughput come una funzione di località temporale e spaziale

La Fig. 1.3 è una rappresentazione grafica che caratterizza il *throughput* di lettura di un sistema di memoria (come un Intel i7 Haswell) in funzione della località temporale e della

località spaziale. La montagna mostra diverse "**creste**" (ridges) che corrispondono alle regioni dove il *working set* si adatta interamente nelle cache L1, L2, L3 e nella memoria principale, rivelando differenze di velocità di accesso di oltre un ordine di grandezza. I "pendii" (slopes) sulla montagna illustrano come il *throughput* diminuisce all'aumentare dello *stride*, evidenziando l'importanza della località spaziale anche quando il *working set* è troppo grande per le cache. [3]

La **persistenza** si riferisce alla capacità di una memoria di mantenere i dati senza alimentazione elettrica.

Come discusso in precedenza la gerarchia di memoria è una soluzione economica per fornire una grande **capacità** di memoria ad alta velocità, sfruttando il principio di località affrontato nei paragrafi antecedenti. Affronteremo la persistenza delle memorie in relazione alla loro volatilità/non volatilità analizzandone anche il **costo per bit** e la capacità, sia a livello di singolo chip/dispositivo, sia a livello di sistema integrato, nei capitoli successivi, mostrando come questi fattori influenzino la progettazione complessiva delle architetture dei computer.

Capitolo 2

Memorie volatili vs Memorie non volatili

2.1 Memorie volatili (RAM, DRAM, SRAM): caratteristiche e utilizzi

Le memorie volatili come già spiegato brevemente in precedenza sono quelle che perdono i dati quando l'alimentazione viene interrotta. Le principali tipologie di memoria volatile nel contesto dei computer sono la DRAM (Dynamic Random-Access Memory) e la SRAM (Static Random-Access Memory). Entrambe sono forme di RAM, che è un termine generico per il tipo di memoria che può essere letta e scritta in qualsiasi ordine.

1. **SRAM** (Static Random-Access Memory)

La SRAM è il tipo di memoria più veloce disponibile e viene utilizzata principalmente per la cache della CPU. Utilizza tipicamente sei transistor[1] per bit per mantenere l'informazione, senza la necessità di essere riaggiornata costantemente, da cui la "S" in SRAM (statico) che sottolinea ciò. È organizzata come una matrice di celle 6T vedi Fig. 2.1.

-**Latenza** (velocità di accesso): I tempi di accesso per i registri variano da 0,1 a 0,2 ns, mentre le cache di primo livello (L1) hanno latenze di circa 1ns mentre quelle di terzo livello (L3) sono tipicamente dalle 2 alle 8 volte più lente delle L2, ma comunque almeno 5 volte più veloci di un accesso alla DRAM.

-**Capacità**: Capacità molto inferiore rispetto alla DRAM o alle memorie di massa a causa del suo costo e della sua complessità. Le cache on-chip[2] variano da 4KiB (per i registri) a 60MiB o più per le cache L3.

-**Costo e Consumo energetico**: È la più costosa per bit, ma richiede un consumo energetico minimo per mantenere la carica in modalità standby. Nella Fig. 2.2 si può notare il confronto in termini di consumo energetico tra la SRAM e la DRAM di cui ora tratteremo.

[1] Un transistor è un componente elettronico che funziona come interruttore o amplificatore dei segnali elettrici. È il blocco fondamentale nei circuiti digitali e analogici, utilizzato nei microprocessori, nelle memorie e in molti altri dispositivi elettronici

[2] Memorie cache integrate direttamente all'interno del chip del processore (CPU)

-**Utilizzi**: Impiegata principalmente per i registri della CPU e per tutti i livelli di cache (L1, L2, L3), spesso integrati direttamente sul chip del processore.

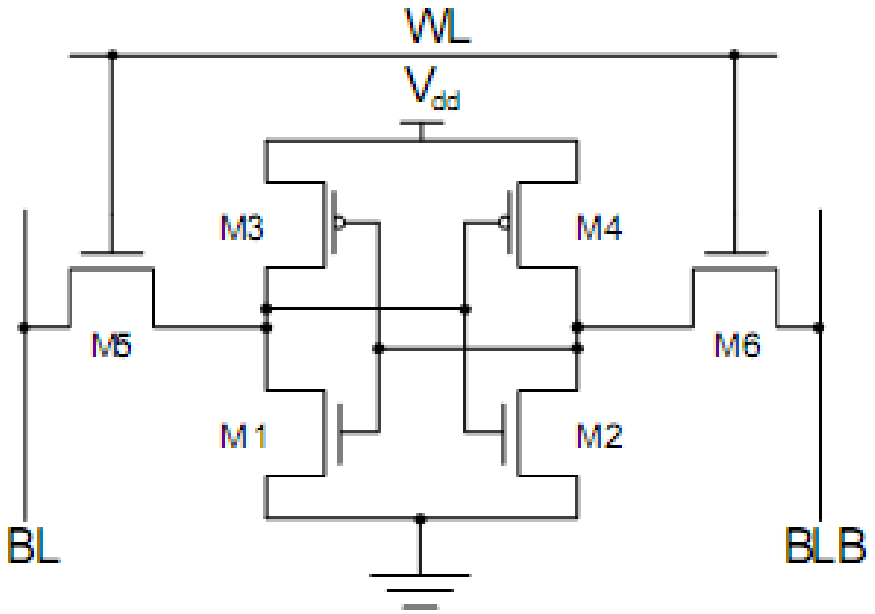


Figura 2.1: Cella SRAM base a 6 transistor

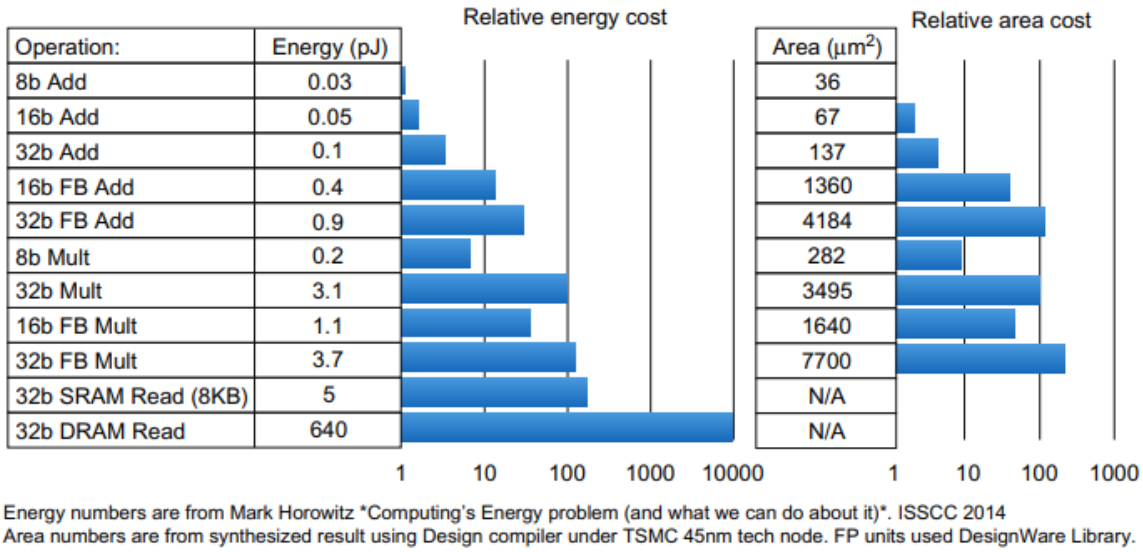
2. **DRAM** (Dynamic Random-Access Memory)

La DRAM è la tecnologia alla base della memoria principale. La crescita delle DRAM è rallentata drasticamente rispetto al passato dove quadruplicava ogni 3 anni. Utilizza un singolo transistor (che agisce come condensatore) per bit e la cella viene comunemente chiamata 1T-1C (1 transistor e 1 condensatore) vedi Fig. 2.3, il che le consente di essere molto più densa (e quindi più economica per bit) della SRAM. La "D" in DRAM sta per "dinamica", perché le informazioni immagazzinate devono essere periodicamente aggiornate per evitare la perdita di dati dovuta alla dispersione della carica del condensatore. Questo refresh rende la memoria occasionalmente non disponibile per gli accessi. Le DRAM sono spesso organizzate in banchi multipli per aumentare la larghezza di banda e migliorare la gestione dell'energia, consentendo accessi sovrapposti o interlacciati.

-**Latenza**: È significativamente più lenta della SRAM. I tempi di accesso per la memoria principale variano tipicamente da 30 a 150 ns, con valori tipici di 50-100 ns negli attuali sistemi.

-**Capacità**: Possiede una capacità molto più elevata della SRAM. La memoria principale varia da 1-2 GB (dispositivi mobili personali) a 8-64 GB (Laptop/Desktop) tralasciando i server di fascia alta con valori fino 256 GB

-**Costo e Consumo energetico**: È un compromesso tra velocità e costo, il consumo energetico include sia potenza dinamica (durante letture/scritture) che quella statica (standby).



Operation:	Energy (pJ)	Area (μm²)
8b Add	0.03	36
16b Add	0.05	67
32b Add	0.1	137
16b FB Add	0.4	1360
32b FB Add	0.9	4184
8b Mult	0.2	282
32b Mult	3.1	3495
16b FB Mult	1.1	1640
32b FB Mult	3.7	7700
32b SRAM Read (8KB)	5	N/A
32b DRAM Read	640	N/A

Energy numbers are from Mark Horowitz *Computing's Energy problem (and what we can do about it)*. ISSCC 2014
Area numbers are from synthesized result using Design compiler under TSMC 45nm tech node. FP units used DesignWare Library.

Figura 2.2: Confronto del consumo energetico e dell'area occupate sul chip (die area) delle operazioni aritmetiche, e del costo energetico degli accessi alla SRAM e alla DRAM.

Come mostrato in Fig. 2.2, se prendiamo come esempio un'addizione floating-point a 32-bit notiamo che utilizza 30 volte l'energia di un'addizione tra interi a 8-bit, con la differenza d'area che è anche più grande (circa 60 volte). Tuttavia la differenza maggiore è nella memoria, infatti un accesso a 32-bit alla DRAM consuma 20.000 volte più energia rispetto ad una addizione a 8-bit. Mentre una piccola SRAM è 125 volte più efficiente in termini di energia rispetto alla DRAM.

-**Utilizzi**: Memoria principale nei computer (Desktop e Laptop) , dove vengono temporaneamente caricati i programmi e i dati in uso, per consentire al processore un accesso rapido.

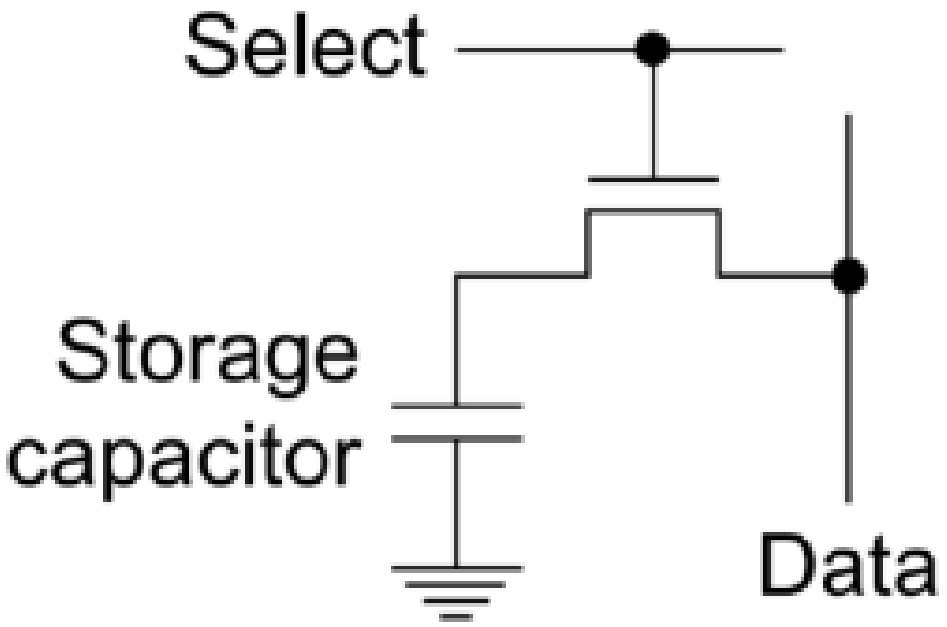


Figura 2.3: Cella DRAM

2.2 Memorie non volatili (ROM, EEPROM, FLASH, PCM): caratteristiche e utilizzi

Le memorie non volatili sono quelle che mantengono i dati anche in assenza di alimentazione. Le principali memorie non volatili sono: Le ROM, di cui abbiamo diverse tipologie come le PROM (ROM programmabile una sola volta), EPROM (erasable ROM) ed EEPROM (riscrivibile elettricamente) più versatile, ma anche più lenta e costosa.

1. ROM (Read-Only-Memory)

La ROM è una memoria non volatile che contiene dati permanenti, scritti una volta durante la produzione e non modificabili dall'utente. Viene generalmente utilizzata per conservare firmware o istruzioni essenziali per l'avvio del sistema. Essendo a sola lettura è molto sicura e stabile nel tempo.

2. EEPROM (Electrically Erasable Programmable Read-Only Memory)

L'EEPROM è una memoria non volatile che può essere letta, cancellata e riscritta elettricamente a livello di singolo byte. È usata in applicazioni dove è necessario aggiornare i dati in modo persistente, come nelle configurazioni hardware. Essa è più lenta della Flash, e possiede cicli di scrittura limitati ma pur sempre superiori alla ROM.

3. FLASH

La memoria Flash [4][5] è una memoria a semiconduttore non volatile, un tipo di EEPROM, sebbene sia normalmente "solo lettura", può essere cancellata e riscritta. Il tipo Flash più denso e adatto per memorie non volatili su larga scala è la NAND Flash che ha come punto a sfavore l'accesso sequenziale e la scrittura più lenta. La memoria Flash è il dispositivo di archiviazione standard nei dispositivi mobili personali (PDM). La cella base di una memoria Flash è un transistor a porta flottante (Floating-Gate Transistor o FGT) che può trattenere elettroni per rappresentare lo stato logico. Vedi Fig. 2.4

-**Funzionamento**: Le operazioni di lettura sulla Flash sono sequenziali come detto in precedenza e leggono un'intera pagina, che può essere di 512 byte, 2KiB o 4KiB. La scrittura richiede che la memoria sia cancellata prima di essere sovrascritta, e la cancellazione avviene in blocchi, non in singoli byte o word. Se si volessero modificare alcuni dati di un blocco, bisognerebbe prima leggerlo tutto in memoria temporanea, sostituire i dati che si vogliono cambiare con i nuovi valori per poi scrivere di nuovo l'intero blocco sulla Flash, cancellando prima quello vecchio.

-**Latenza e Velocità**: Ha un lungo ritardo per l'accesso al primo byte da un indirizzo casuale circa 25 μs ma può fornire il resto di un blocco di pagina a circa 40 MiB/s. Come paragone prendendo una DDR4 SDRAM impiega circa 40 ns per l'accesso al primo byte e può trasferire il resto della riga a 4,8GiB/s. Tuttavia, se confrontiamo i tempi per trasferire 2KiB, la NAND Flash impiega circa 75 μs mentre la DDR SDRAM meno di 500 ns. Ciò rende la Flash 150 volte più lenta e ci fa capire come essa non sia la migliore candidata per rimpiazzare la memoria principale, ma una buona candidata per rimpiazzare il disco magnetico essendo da 300 a 500 volte più veloce.

-**Capacità e Costo**: La capacità per chip Flash è aumentata rapidamente negli ultimi anni di circa 50%-60% all'anno, raddoppiando circa ogni 2 anni. Nel 2011, il costo per GiB della Flash era di circa $2/GiB, rendendola molto più economica per bit rispetto alla DRAM, ma 15-25 volte più costosa rispetto ai dischi magnetici.

-**Consumo Energetico**: Richiede un consumo significativamente inferiore quando non è in lettura o scrittura.

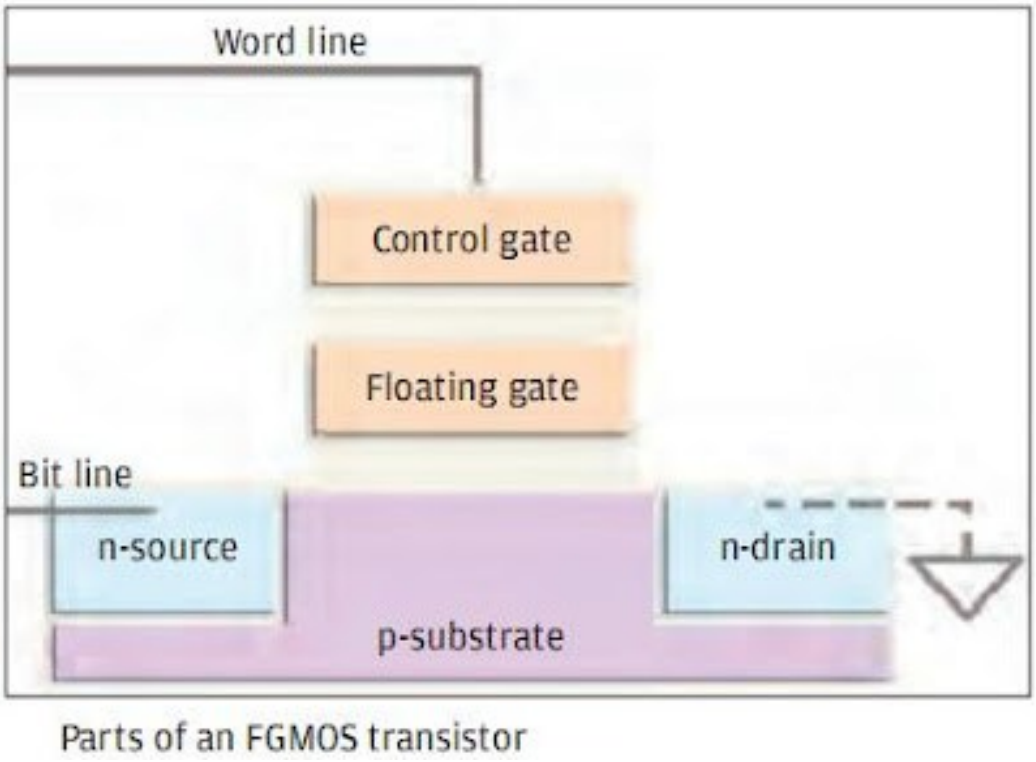


Figura 2.4: Transistor a porta flottante (FTG)

4. PCM (Phase-Change Memory)

Le PCM [6] o meglio Memorie a cambiamento di fase, sono considerate una delle principali tecnologie emergenti nel campo delle memorie non volatili. Pur non essendo recentissimo come concetto, lo sono per quanto riguarda l'applicazione commerciale e la maturazione tecnologica. Sono attivamente sviluppate per sostituire o affiancare le Flash e DRAM grazie alle lore caratteristiche ibride: non volatilità, velocità vicina alla DRAM e buona resistenza.

-**Funzionamento**: La tecnologia si basa su un piccolo elemento riscaldante che cambia lo stato di un substrato tra la sua forma cristallina e amorfa, che posseggono proprietà resistive diverse. La lettura avviene rilevando la resistenza, la scrittura applicando una corrente per cambiare la fase del materiale.

L'argomento verrà successivamente approfondito trattando una emergente applicazione delle PCM in ambito di Sinapsi Artificiali (Artificial Synapses)

2.3 Confronto tecnico e funzionale tra VM e NVM

Le memorie volatili e non volatili si differenziano per una serie di caratteristiche tecniche e funzionali che influenzano profondamente la loro applicazione nei sistemi informatici. Il confronto tra queste due categorie è fondamentale per comprendere le scelte progettuali nell'architettura dei computer, nei dispositivi embedded[3] e nei sistemi di storage.

Una prima distinzione riguarda la **persistenza** dei dati: le memorie volatili, come DRAM e SRAM, perdono il contenuto memorizzato in assenza di alimentazione elettrica, rendendole inadatte allo storage permanente. Le memorie non volatili come EEPROM e Flash, mantengono invece i dati anche dopo lo spegnimento del dispositivo.

Dal punto di vista della **latenza**, le memorie volatili offrono tempi di accesso estremamente ridotti, misurabili nell'ordine dei nanosecondi. Le DRAM costituiscono la memoria principale di un sistema proprio per la loro velocità, inferiore comunque a quella delle SRAM usata nella cache. Le memorie non volatili invece hanno latenze sensibilmente più elevate, talvolta nell'ordine dei microsecondi o millisecondi, rendendole ideali per funzioni di archiviazione.

Anche la **durabilità** e il **numero di cicli** di scrittura rappresentano un parametro chiave, infatti, le memorie volatili non soffrono di limiti significativi in termini di scritture mentre molte memorie non volatili presentano un'usura progressiva con un numero massimo di cicli di programmazione.

Un altro elemento molto importante è **l'efficienza energetica**. Le memorie volatili richiedono alimentazione costante per mantenere lo stato dei dati, ciò contribuisce non poco al consumo energetico. Al contrario le memorie non volatili non consumano energia in stato di inattività riducendo così il consumo di energia.

Il concetto di gerarchia della memoria sfrutta queste differenze, posizionando le memorie più veloci e costose come SRAM e DRAM più vicine alla CPU e quelle più lente e capaci come Flash e HDD più distanti, in una configurazione progettata per ottimizzare il compromesso tra velocità, capacità e costo.

[3] Dispositivi progettati per fare una cosa sola, ma in modo efficiente, affidabile e continuo

Capitolo 3

Evoluzione delle memorie non volatili

3.1 RRAM and MRAM

3.1.1 RRAM

Le **RRAM** sono una classe di tecnologie di memoria emergenti che si basano sul principio di distinguere il contenuto della memoria in base alla resistenza della cella di bit. L'idea di utilizzare memorie basate sulla resistenza viene avvalorata dal fatto che una varietà di ossidi presenta una commutazione resistiva (**Resistive Switching**[4]), ovvero la tensione varia in funzione della tensione applicata tra di essi e questi ossidi metallici binari sono facilmente compatibili con CMOS. Ci sono differenti tipi di RRAM come quella della RRAM a ponte conduttivo (CBRAM) che fa uso di uno ione metallico per la formazione del filamento e si basa sul movimento degli ioni metallici per determinare la resistenza del dispositivo e la successiva commutazione, e della RRAM a ossido (OxRRAM) (vedi Fig. 3.1) di cui tratteremo in seguito mediante uno scenario applicativo.

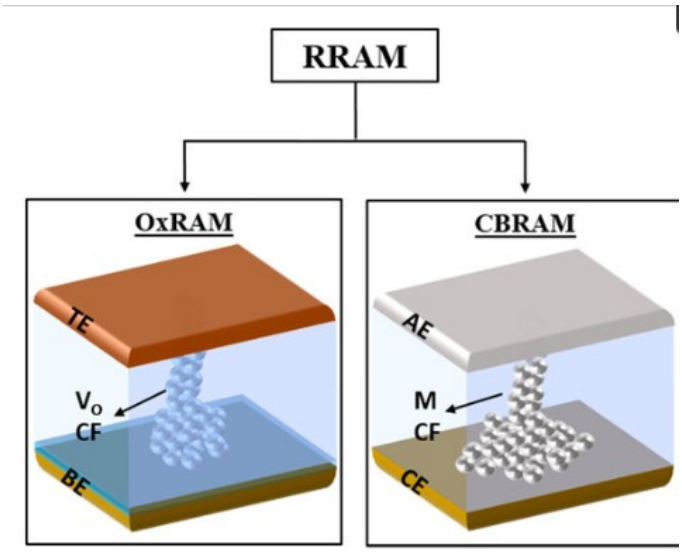


Figura 3.1: Rappresentazione schematica di classificazione delle RRAM sulla base dei meccanismi di commutazione, OxRRAM e CBRAM.

[4] Capacità di alcuni materiali, in particolare ossidi di metalli di transizione, di modificare in modo reversibile il proprio stato di resistenza elettrica quando vengono applicate opportune tensioni.

Esistono due tipi di RRAM, ovvero la RRAM unipolare e quella bipolare, distinte in base alle tensioni necessarie per eseguire la commutazione. È importante comprendere le caratteristiche del dispositivo per comprendere le applicazioni e i compromessi della *bitcell*.

Nel caso della **commutazione bipolare**, che è l'uso più comune delle RRAM, essa comporterebbe l'applicazione di tensioni positive tra l'elettrodo superiore e quello inferiore per eseguire l'operazione di "set" e l'applicazione di tensioni negative tra l'elettrodo superiore e quello inferiore per eseguire l'operazione di "reset".

Nel caso della **commutazione unipolare**, solo una tensione positiva è sufficiente per subire cambiamenti set-reset e viceversa. Vedi Fig. 3.3.

In Fig. 3.2 si possono notare le caratteristiche IV (Corrente-Tensione) per una RRAM a commutazione bipolare, in cui la tensione di set è elevata nell'intervallo di 3-3.2 V e quella di reset è nell'intervallo -ve 3-3.2 V. [8]

La curva I-V descrive come la corrente attraverso il dispositivo cambia in risposta a una tensione applicata. Essa mostra le due modalità operative ovvero *set* e *reset* e il comportamento non lineare e bistabile del dispositivo, infatti, può trovarsi in due stati resistivi stabili e commutare tra essi sotto certe condizioni di tensione.

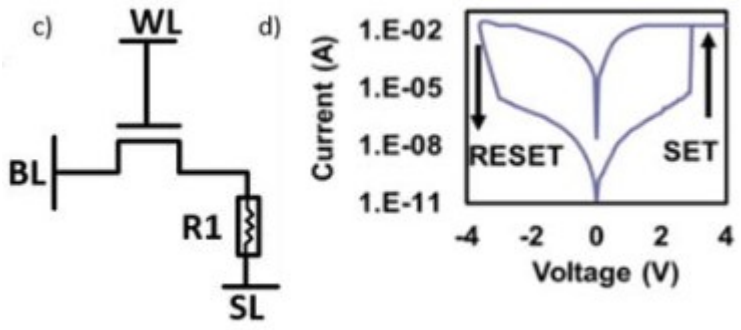


Figura 3.2: RRAM bitcell a sinistra e RRAM caratteristiche I-V a destra

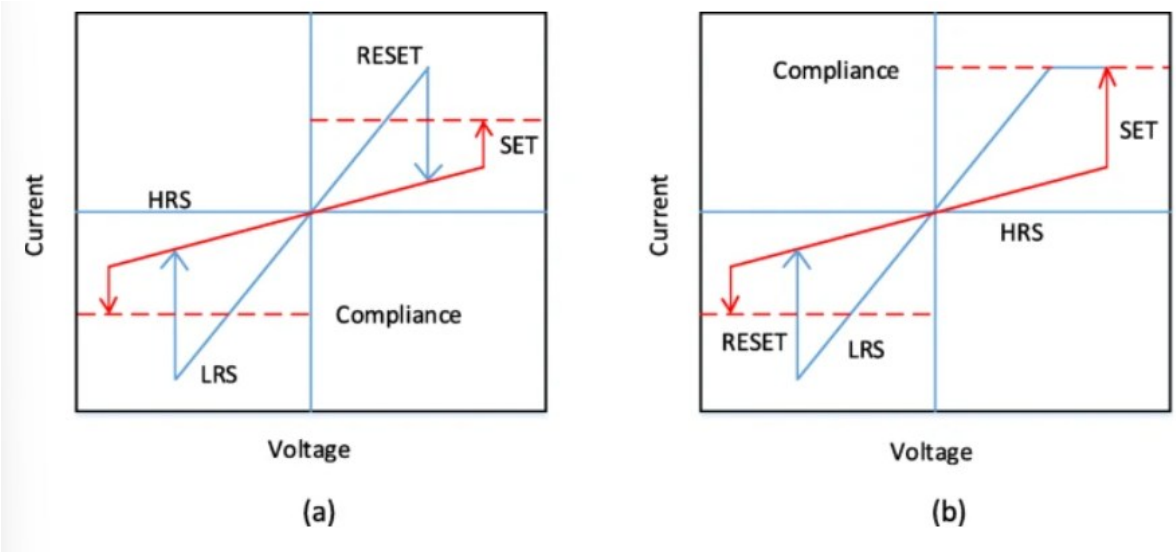


Figura 3.3: a) Commutazione unipolare, b) Commutazione bipolare

3.1.2 MRAM

La MRAM come la RRAM appartiene anch'essa ad una classe di memorie che si basa sulla resistenza di diversi stati per memorizzare contenuti diversi sulla cella di bit e sul costruire concetti di magnetismo/effetto Hall di spin per modulare la resistenza della cella di bit, da cui il nome di Memoria Magneto-resistiva. Essa è una memoria non volatile basata sullo spin (Spintronica[5]), utilizza la carica magnetica per memorizzare dati al posto della carica elettrica degli elettroni. Il principio alla base della MRAM è la magneto-resistenza.

Le varie tipologie di MRAM sono: MRAM anisotropica (AMR-MRAM), MRAM a valvola di spin (SV-MRAM), MRAM a pseudo valvola di spin (PSV-MRAM), MRAM a giunzione magnetica di tunnel (MTJ) e MRAM a effetto Hall (SOT-MRAM).

Quella comunemente utilizzata si basa su MTJ, che include uno strato libero, uno strato di ossido (che funge da barriera tunnel) e strati fissi vedi Fig. 3.4. Applicando un campo magnetico esterno, la magnetizzazione dello strato libero può essere variata. I dati vengono memorizzati in base alla direzione della magnetizzazione dello strato libero, in relazione alla magnetizzazione dello strato fisso. [7]

[5] La spintronica è l'elettronica basata sullo spin, i dispositivi spintronici non richiedono corrente elettrica per mantenere il loro *spin*.

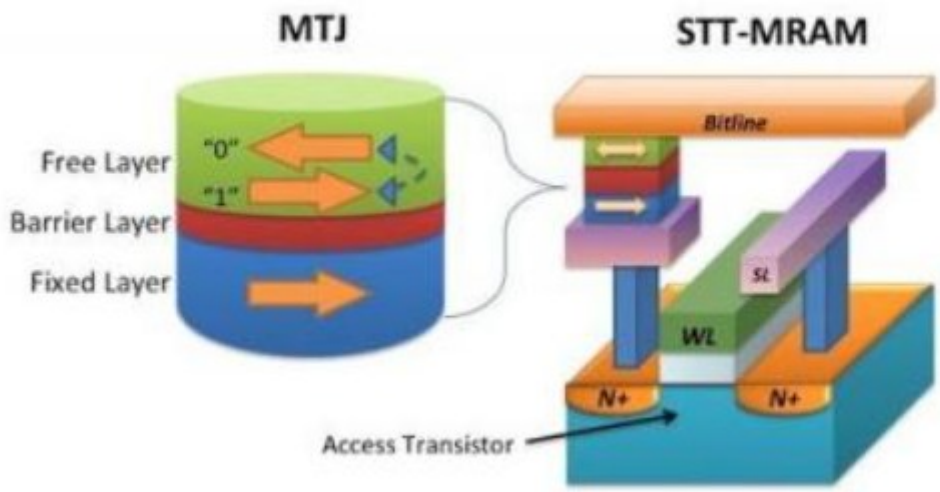


Figura 3.4: Layer differenti in una MTJ a sinistra, SST-MRAM a destra

Una soluzione avanzata è la STT-MRAM (Spin Transfer Torque MRAM) vedi Fig. 3.4, dove l'orientamento magnetico viene stabilito utilizzando una corrente di elettroni polarizzati, mentre nella MRAM tradizionale si utilizza un campo magnetico esterno. Per la lettura dei dati viene applicata una piccola tensione per far fluire la corrente attraverso la MTJ. In questo modo, la resistenza della MTJ può essere misurata dalla corrente. Viene poi confrontata questa corrente con una di riferimento per leggere i dati. [7]

Il vantaggio della bitcell paragonato a quella della RRAM è l'estremamente alta resistenza di quest'ultima (nell'ordine di circa 10^{15} cicli) e con il voltaggio e la latenza per la scrittura rispettivamente: il primo leggermente più basso e il secondo più basso rispetto alla RRAM.

Similmente al caso delle RRAM, la direzione della corrente determina la commutazione della MRAM e il flusso di corrente dallo strato fisso allo strato libero è responsabile della commutazione dello strato libero dallo stato parallelo (bassa resistenza "1") a quello antiparallelo (alta resistenza "0"). [8]

3.2 Phase Change Memory (PCM) e OxRAM (Oxide-based Resistive RAM)

3.2.1 Phase Change Memory

La memoria a cambiamento di fase o PCM utilizza la proprietà unica del vetro calcogenuro[6] che possiede due stati di materia: amorfo o cristallino. Essa sfrutta la differenza di resistenza per memorizzare le informazioni in bit. La resistività varia quando i materiali a cambiamento di fase commutano tra due fasi quando viene applicato calore mediante impulsi elettrici: alta per lo stato **amorfo** e bassa resistività per quello **cristallino**. Il materiale calcolgenuro solitamente utilizzato è il Ge_2-Sb_2-Te_5 o GST.

All'interno di una cella di una PCM si trova un materiale a cambiamento di fase, insieme ad un elemento riscaldante inserito tra 2 elettrodi vedi Fig. 3.5.

L'iniezione di corrente tra l'interfaccia del PCM e l'elemento riscaldante porta il calcogenuro a una temperatura superiore rispetto al suo punto di cristallizzazione, ma inferiore al punto di fusione, rendendo il materiale cristallino, questa operazione è chiamata operazione di "SET." In alternativa applicando un'alta tensione al materiale cristallino, la conduttività provoca la deviazione della corrente e il ritorno al suo stato amorfo, operazione nota come operazione di "RESET."

Durante l'operazione di lettura viene applicata una bassa potenza, rivelandone così la resistività. [7]

Per l'operazione di ***set*** viene applicata una corrente moderata per un tempo più lungo (alcune decine di ns), il materiale così si riscalda fino alle temperatura di cristallizzazione (~200 °C) che poi si raffredda lentamente, ciò induce la formazione di una struttura cristallina con bassa resistenza che equivale ad un 1 logico.

Per l'operazione di ***reset*** viene applicata una corrente più intensa ma per un tempo molto breve, il materiale fonde e arriva a ~600°C con conseguente raffreddamento molto rapido, non lasciando tempo agli atomi di ordinarsi, si ottiene così una struttura amorfa con alta resistenza che equivale ad uno 0 logico.

[6] I vetri calcogenuri sono amorfi, vengono usati per la Phase Change Memory (PCM) grazie alla loro capacità di passare rapidamente tra stato amorfo e cristallino

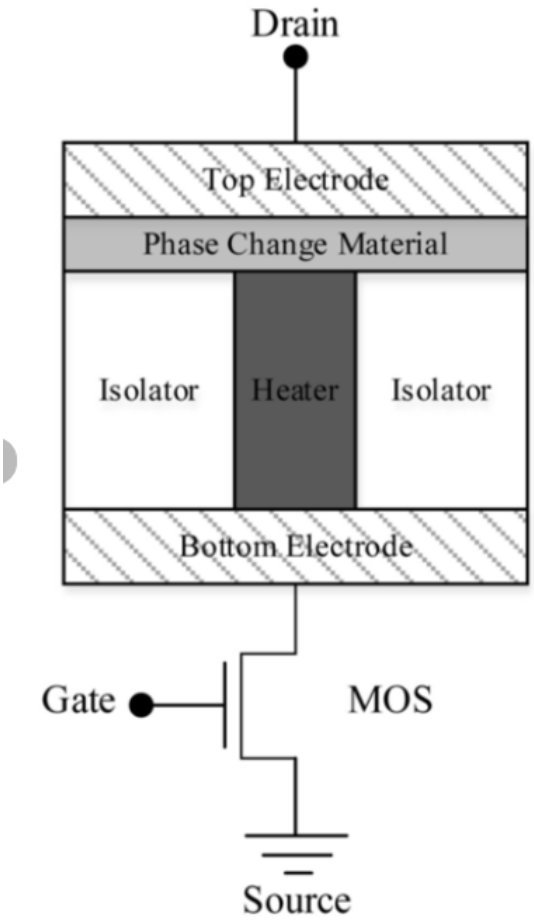


Figura 3.5: Struttura di una cella PCM

3.2.2 OxRRAM

Modalità di commutazione della resistenza

Una memoria ad accesso casuale resistiva o più comunemente RRAM è costituita da una cella di memoria a commutazione resistiva, brevemente accennata in precedenza (vedi Sezione 3.1.1) con una struttura metallo-isolante-metallo, generalmente nota come struttura MIM. La struttura è costituita da uno strato isolante inserito tra due elettrodi metallici vedi Fig. 3.6.

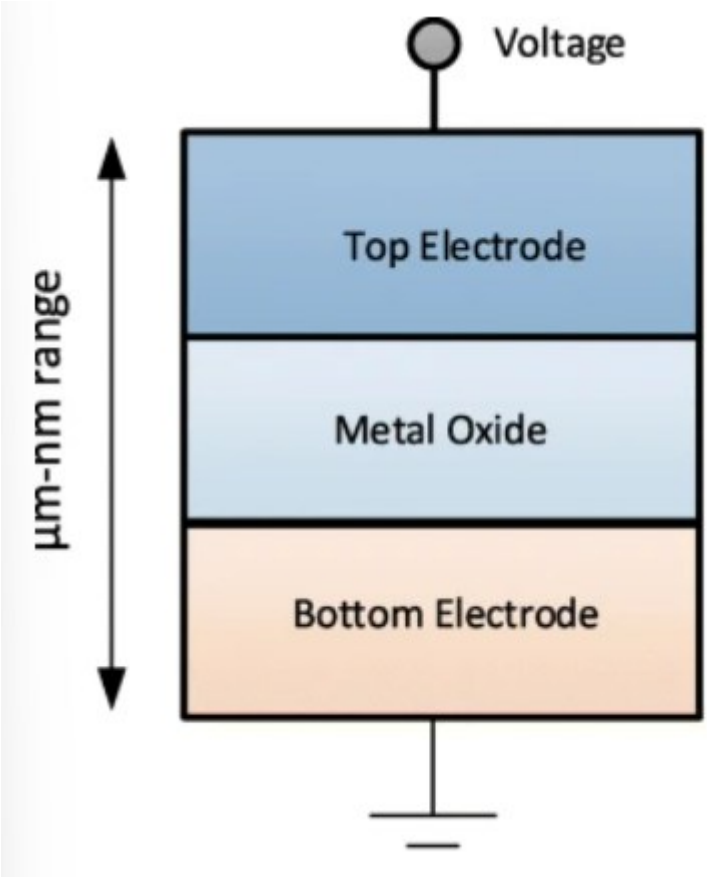


Figura 3.6: Schema della struttura metallo-isolante-metallo per RRAM

Una RRAM così preparata si trova inizialmente nelle stato di alta resistenza (HRS), per commutare il dispositivo dall'HRS all'LRS (Bassa resistenza), l'applicazione dell'impulso di alta tensione consente la formazione di percorsi conduttivi nello strato di commutazione e la cella RRAM viene commutata in un LRS. Questo processo, che si verifica a causa della rottura morbida della struttura metallo-isolante (MIM), è solitamente indicato come **elettroformatura** e la tensione a cui avviene questo processo è indicata come tensione di formatura (V_f).

Meccanismo di commutazione resistivo

La commutazione della cella RRAM si basa sulla crescita del filamento conduttivo all'interno di un dielettrico. Il filamento è un canale con un diametro molto piccolo, dell'ordine dei nanometri, che collega gli elettrodi superiore e inferiore della cella di memoria. Lo stato di bassa resistenza (LRS) e alta resistenza (HRS) nominati in precedenza si ottengono rispettivamente quando il filamento è connesso per il primo e disconnesso per il secondo.

In base alla composizione del filamento conduttivo , le RRAM possono essere classificate in due tipi: CBRAM basate su ioni metallici e le OxRRAM basate su filamenti di ossigeno vacanti[7].

Nelle RRAM basata su vacanze di ossigeno, il meccanismo fisico responsabile della commutazione resistiva è generalmente associato alla generazione di vacanze di ossigeno. Una volta che si verifica la rottura morbida del dielettrico, gli atomi di ossigeno vengono espulsi dal reticolo applicando un campo elettrico elevato verso l'interfaccia dell'anodo e diventano ioni di ossigeno O^{2-} mentre le vacanze di ossigeno vengono lasciate nello strato ossido. Se i metalli nobili vengono utilizzati come materiali per l'anodo al fine di formare uno strato di ossido interfacciale, gli ioni di ossigeno reagiscono con i materiali dell'anodo e vengono scaricati come ossigeno neutro non reticolare. Pertanto, l'interfaccia elettrodo/ossido si comporta come un serbatoio di ossigeno. Successivamente, l'accumulo delle vacanze di ossigeno nell'ossido di massa, commuta la cella RRAM allo stato a bassa resistenza (LRS) quando si forma il filamento conduttivo e la corrente scorre nel dispositivo. Per riportare il dispositivo allo stato di alta resistenza, si verifica il processo di *reset,* durante il quale gli ioni di ossigeno O^{2-} migrano di nuovo all'ossido di massa dall'interfaccia anodica e si combinano con le vacanze di ossigeno per ossidare i precipitati metallici del filamento conduttore e quindi romperlo parzialmente, riportando la cella RRAM allo stato HRS. [9]

[7] Un difetto puntuale in un reticolo cristallino in cui un sito normalmente occupato da un atomo di ossigeno è vuoto

3.3 PCM e OxRRAM come sinapsi nelle reti neurali spiking

Le memorie a cambiamento di fase (PCM) e le memorie a ossido metallico (OxRRAM) possono essere utilizzate come elementi sinaptici grazie alla conduttività regolabile, alla compatibilità con i processi di fabbricazione CMOS avanzati e alla scalabilità.

In questa sezione tratteremo un'applicazione di queste memorie non volatili emergenti, nell'ambito dei recenti progressi nell'implementazione della plasticità sinaptica, mettendo in luce alcune proprietà di questi dispositivi, solitamente considerate non ideali ma che possono migliorare le prestazioni delle reti neurali Spiking[8] (SNN) addestrate con un algoritmo di apprendimento non supervisionato denominato Spike-Timing Dependent Plasticity[9] (STDP).

Le memorie resistive RRAM sono candidate estremamente interessanti per l'implementazione di sinapsi plastiche, la OxRAM può essere programmata consumando poca energia e può essere integrata semplicemente con tecnologie CMOS avanzate. Tuttavia, questi dispositivi soffrono di un'elevata variabilità della conduttanza[10] da ciclo a ciclo e da dispositivo a dispositivo, presentando una sfida considerevole.

Un altro candidato promettente è la memoria a cambiamento di fase (PCM) in quanto può essere usata come memoria analogica, dovuto al fatto che possiede la capacità di modulare la resistenza elettrica in modo continuo, non solo tra due stati binari, ma anche su più livelli intermedi.

3.3.1 OxRAM come sinapsi artificiale

I dispositivi OxRAM presi in esame sono composti da una pila Tin/HfO_2/Ti/TiN, dove entrambi gli strati di HfO_2 e Ti hanno uno spessore di 10 nm. Ogni cella OxRAM è collegata ad un transistor NMOS (1T1R) vedi Fig. 3.7(a), che seleziona la cella e controlla la corrente durante le operazioni di *set* e *forming*. Sono state collegate in parallelo più celle OxRAM e ognuna di queste celle può trovarsi in HCS (Stato di alta conduttanza) o LCS (Stato di bassa conduttanza). Applicando la regola di apprendimento STDP (Spike-Timing Dependent Plasticity) se il neurone presinaptico si attiva appena prima di quello postsinaptico, si verifica

[8] Le reti neurali Spiking (SNN) sono un tipo avanzato di rete neurale artificiale che mima più da vicino il funzionamento del cervello biologico.
[9] La regola di apprendimento STDP è una delle regole fondamentali di apprendimento sinaptico usate nei modelli di reti neurali spiking ed ispirata dal funzionamento del cervello biologico.
[10] La variabilità di conduttanza si riferisce al comportamento fluttuante o non costante della conduttanza elettrica di un dispositivo di memoria.

un evento di **potenziamento** a lungo termine (LTP) e ciascuna cella OxRAM della sinapsi ha una probabilità P_{LTP} di passare all'HCS.

Se viceversa il neurone postsinaptico si attiva prima si ha un evento di **depressione** con probabilità P_{LTD} di passare all'LCS. In questo modo, le sinapsi simulano l'apprendimento biologico.

I due stati conduttivi HCS e LCS corrispondono rispettivamente all'1 e allo 0 logico, per funzionare bene come memoria binaria le due distribuzioni, vedi Fig. 3.7(b), non devono sovrapporsi. Si valuta perciò questa separazione tramite il **Memory Window** a 3σ ($MW_{3\sigma}$).

$$MW_{3\sigma} = HCS_{-3\sigma} / LCS_{+3\sigma}$$

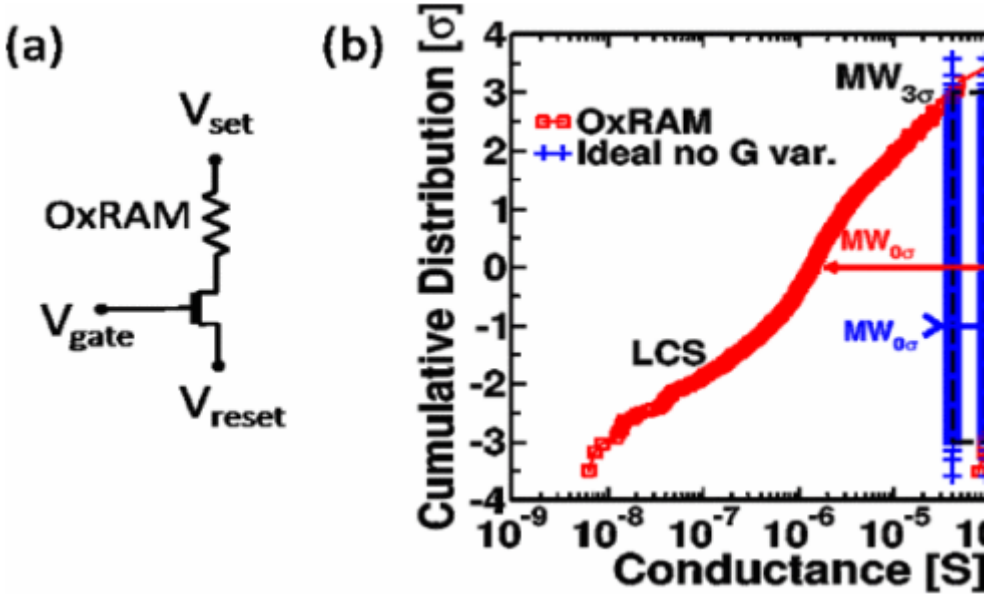


Figura 3.7: a)Configurazione della cella 1T1R, b)Distribuzione cumulativa di LCS e HCS

Esperimento

L'esperimento consiste in una rete neurale ad un solo *layer*, progettata per rilevare le auto su un'autostrada ripresa con un sensore di immagini basato su eventi.

In primo luogo, è stato studiato l'impatto del numero di livelli sinaptici e della finestra di memoria OxRAM sulle prestazioni di rete. Il numero n di OxRAM per sinapsi è stato modificato per variare il numero dei livelli. In Fig. 3.8 si può notare il punteggio di rilevamento in funzione della finestra di memoria 3σ per diversi numeri di livelli sinaptici. Da ciò si evince che l'aumento dei livelli sinaptici ovvero un maggior numero di celle non migliora le prestazioni della SNN: una sinapsi con due livelli infatti è sufficiente per

questo tipo di applicazioni, mentre aumentare la finestra di memoria sì, il tasso di rilevamento cresce, fino a saturarsi a ≈ 0.96.

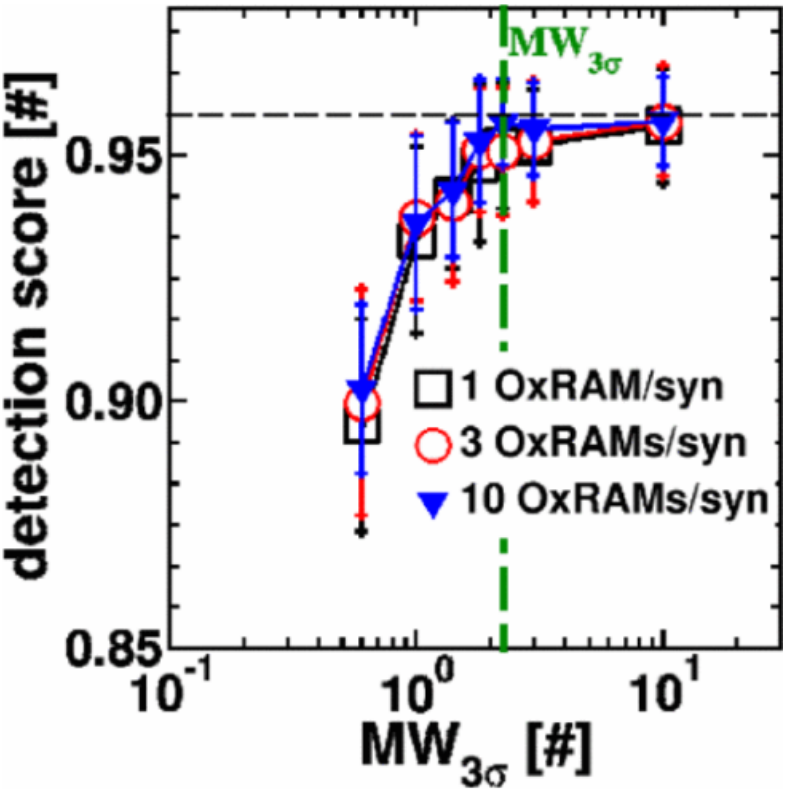


Figura 3.8: Punteggio di rilevamento in funzione della finestra di memoria 3σ per diversi numeri di OxRAM per sinapsi

In secondo luogo, è stato analizzato l'impatto della variabilità della conduttanza. Sono stati confrontati due casi: le prestazioni dell'applicazione vista poco fa e un dispositivo artificiale a variabilità zero mantenendo sempre la stessa $MW_{3\sigma}$. In Fig. 3.7 (b) si nota come il punteggio di rilevamento sia 0.63 per la sinapsi artificiale senza variabilità e 0.952 per la cella OxRAM reale. La variabilità nella conduttanza aiuta la rete neurale a trovare una gamma più ampia di pesi sinaptici, migliorando così l'apprendimento. [10]

3.3.2 PCM come sinapsi artificiale bidirezionale

Per questo esperimento è stato utilizzato un materiale a cambiamento di fase basato su GST integrato in dispositivi PCM che comprendono un elemento di accumulo a parete (l'elettrodo inferiore) su una piattaforma tecnologica CMOS a 130 nm vedi Fig. 3.9. L'elettrodo inferiore ha uno spessore nominale di 4 nm e una larghezza di 50nm.

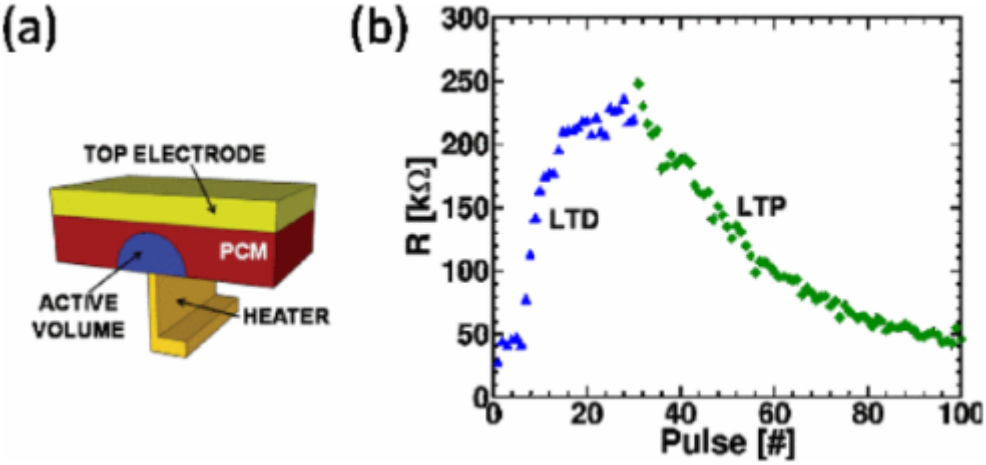


Figura 3.9: a)Disegno schematico dell'elemento di accumulo a parete della PCM, b) Caratteristiche di depressione (LTD) e potenziamento (LTP)

Il riscaldamento avviene partendo dal fondo, evitando il surriscaldamento diretto del materiale a cambiamento di fase (GST). La strategia che viene usata è quella di utilizzare una serie di impulsi identici tutti da 20 ns per ottenere sia una cristallizzazione graduale (potenziamento) che un'amorfizzazione (depressione). Si ha per prima cosa una fase di inizializzazione che avviene con un impulso veloce di 1.25 V che imposta la cella a circa 30 kΩ.

-Depressione graduale: con impulsi da 1.6 V che porta all'aumento della resistenza ossia ad una amorfizzazione graduale.

-Potenziamento graduale: con impulsi da 1.25 V che porta alla diminuzione della resistenza ossia ad una cristallizzazione graduale.

Si evince che la relazione tra la risposta della resistenza e il numero di impulsi applicati non è lineare sia per la amorfizzazione che per la cristallizzazione. Inoltre, è presente anche un'asimmetria tra depressione e potenziamento.

Il modello utilizzato è stato preso in esame per analizzare l'impatto sia della non linearità che dell'asimmetria. Viene usato un modello sinaptico matematico per descrivere la conduttanza G in funzione dei parametri di non linearità (β_+ e β_-):

$$\delta G^{+} = \alpha^{+} \times \exp\left(-\beta^{+} \times \frac{G - Gmin}{Gmax - Gmin} \right)$$

$$\delta G^{-} = \alpha^{-} \times \exp\left(-\beta^{-} \times \frac{Gmax - G}{Gmax - Gmin} \right)$$

dove:

δG^{+} = incremento di peso (potenziamento)

δG^{-} = decremento di peso (depressione)

β = controlla quanto è non lineare la risposta

L'asimmetria dipende dal numero di livelli di potenziamento e depressione, essa migliore quando i numeri di livelli di *set*/*reset* sono bilanciati.

Esperimento

È stato dimostrato che le sinapsi di conduttanza multilivello sono necessarie per ottenere le migliori prestazioni nell'applicazione di classificazione dei caratteri. Dal momento che la PCM consente di ottenere più valori di conduttanza, a differenza della OxRAM, è stata quindi simulata una rete *feed-forward*[11] completamente connessa a un livello per la classificazione MNIST[12]. In Fig. 3.10 possiamo osservare le prestazioni della rete in termini di velocità di classificazione in funzione della linearità della risposta di conduttanza (β). Non solo la **non linearità** della risposta di conduttanza migliora le prestazioni della rete rispetto a quella lineare, ma anche l'aumento della **simmetria** tra il potenziamento e la depressione migliora di molto le *performance* (il simbolo nero in Fig. 3.10).

[11] Una rete feed-forward è il tipo più semplice di rete neurale artificiale, il termine significa che l'informazione scorre solo in avanti, dall'input verso l'output, senza cicli o retroazioni.
[12] MNIST è un dataset standard molto usato per addestrare e testare reti neurali, contiene immagini in bianco e nero di cifre scritte a mano.

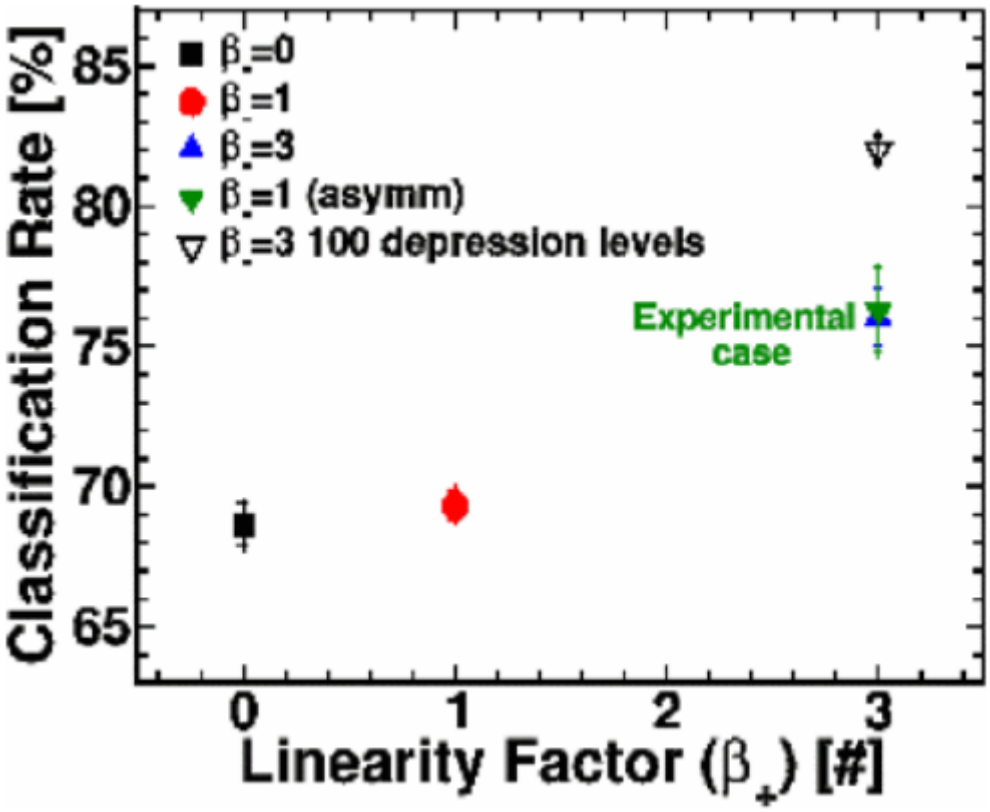


Figura 3.10: Velocità di classificazione in funzione del fattore di linearità

Per illustrare l'impatto della non linearità della risposta di conduttanza durante l'addestramento, possiamo osservare nella Fig. 3.11, l'evoluzione di 15 sinapsi in ~ 20 s. le sinapsi con caratteristiche lineari convergono alla conduttanza minima e massima dopo l'apprendimento, mentre quelle non lineari beneficiano anche dei valori di conduttanza intermedi tra i valori di minimo e massimo. Per il dispositivo sperimentale preso in esame il tasso di classificazione è del 76% invece per 100 livelli di depressione e 200 di potenziamento si ha un tasso dell'82%. [10]

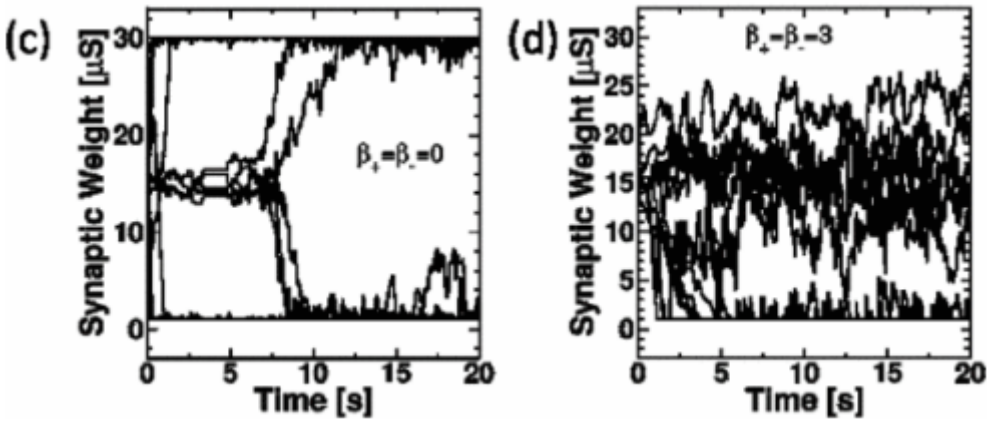


Figura 3.11: c)Evoluzione simulata della conduttanza di 15 sinapsi PCM durante l'apprendimento con risposta lineare, d) non lineare

3.3.4 Conclusioni

Nel contesto dell'elettronica neuromorfica, le memorie resistive non volatili rappresentano una promettente soluzione per la realizzazione di sinapsi artificiali efficienti e scalabili. In particolare, OxRAM e PCM si sono dimostrate due tecnologie chiave grazie alla loro capacità di modulare la conduttanza elettrica in risposta a stimoli esterni, imitando i processi di apprendimento sinaptico del cervello biologico.

Da un lato la OxRAM, basata su ossidi di metallo e vacanze di ossigeno, mostra un comportamento binario (LHS/HCS) ma può essere estesa a più livelli tramite configurazioni in parallelo (multi-cella) e programmazione probabilistica.

Dall'altro la PCM, basata su materiali a cambiamento di fase come GST, offre un controllo più fine della conduttanza grazie alla possibilità di realizzare stati intermedi stabili, sfruttando il processo di amorficazione e cristallizzazione graduale.

Entrambe le tecnologie rappresentano passi concreti verso la realizzazione di hardware neuromorfico efficiente, avvicinando l'elettronica all'elaborazione ispirata al cervello umano.

3.4 STT-RAM come una memoria universale

La STT-RAM (Spin Transfer Torque MRAM) è una soluzione avanzata di MRAM, che si differenzia dalle RAM convenzionali in quanto in essa l'orientamento magnetico viene stabilito utilizzando una corrente di elettroni polarizzati[13], mentre nella MRAM tradizionale si utilizza un campo magnetico esterno.

Per la lettura dei dati, viene applicata una piccola tensione per far fluire la corrente attraverso la MTJ spiegata precedentemente nella sezione dedicata alle MRAM. In questo modo, la resistenza della MTJ basata sulla magnetizzazione può essere misurata dalla corrente. [7]

La STT-RAM ha il potenziale per diventare una memoria universale grazie ai numerosi vantaggi che offre rispetto alle tecnologie di memoria esistenti.

L'articolo che verrà analizzato tratterà del comportamento transitorio di due modelli dinamici di STT-RAM, al fine di valutare la loro accuratezza e applicabilità per la progettazione circuitale e di dispositivi. La STT-RAM è in grado di offrire caratteristiche migliori rispetto alle sue controparti in termini di: non volatilità, densità, consumo energetico, prestazioni, scalabilità ed elevate capacità di integrazione con la tecnologia CMOS.

In Fig. 3.4 si può notare la composizione di una STT-MRAM, che utilizza un MTJ, ovvero un dispositivo a 3 *layer* formato da, un sottile ossido isolante, inserito tra due strati ferromagnetici. La direzione di magnetizzazione di uno strato viene mantenuta fissa o bloccata mentre l'altro strato viene mantenuto libero. Il momento magnetico dello strato libero (FL- *fixed layer*) durante le operazioni di scrittura e i valori logici memorizzati nella MTJ dipendono dalla direzione del FL rispetto allo strato bloccato (PL- *pinned layer*).

Se i due strati ferromagnetici sono antiparalleli tra loro, si ottiene uno stato di alta resistenza (logico '1') e viceversa (logico '0') se sono paralleli tra loro. La differenza tra i due stati di resistenza è data da una metrica nota come rapporto di magnetoresistenza a tunnel (TMR) dato da:

$$\mathrm{TMR} = \frac{\mathrm{R_{AP}} - R_P}{R_P} \times 100$$

Dove R_{AP} rappresenta gli stati di alta resistenza e R_P rappresenta lo stato di bassa resistenza dell'MTJ.

[13] Gli elettroni polarizzati sono elettroni nei quali la maggior parte ha lo spin orientato nella stessa direzione.

Per scrivere nella memoria, è necessaria una corrente maggiore della corrente critica, che causa la variazione del momento magnetico nella FL. Una corrente polarizzata in spin maggiore della corrente critica può fluire da FL a PL. Ciò fa si che gli elettroni, già polarizzati in spin rispetto alla magnetizzazione della PL, attraversino l'ossido fino alla FL, dove esercitano una coppia sugli elettroni, causando una commutazione da parallelo ad antiparallelo, la corrente ora fluisce da PL a FL. Una debole corrente viene quindi fatta passare attraverso la MTJ per rilevare la resistenza tramite un amplificatore di rilevamento per l'operazione di lettura. Maggiore è il TMR, maggiore è la facilità di lettura.

Nella MTJ convenzionale, la direzione del momento magnetico degli strati FM giace nel piano del piano degli strati magnetici (iMTJ[14]), per quest'analisi verranno trattati solo modelli con l'iMTJ.

Modellazione MTJ

Esistono due metodi di modellazione: statico e dinamico. Nella modellazione **statica**, la resistenza dell'MTJ viene impostata in base al verificarsi o meno di un'operazione di commutazione. La condizione di commutazione viene prevista in base al valore della corrente critica (I_C) o dell'angolo critico e al tempo di commutazione. Questi metodi sono spesso implementati utilizzando solo elementi SPICE puri[15].

La modellazione **dinamica**, d'altra parte, include la variazione del vettore di magnetizzazione rispetto al tempo, gli effetti termici e i processi stocastici.

I modelli statici sono più veloci, ma non possono essere utilizzati per osservare la risposta transitoria della memoria poiché un modello statico non descrive come avviene il passaggio di stato, ma solo se e quando è avvenuto. I modelli dinamici eseguono analisi più accurate man mano che si riduce la tecnologia o si progettano dispositivi con velocità più elevate.

Analizziamo due modelli dinamici:

1) Modello di Zihan Xu: basato sull'equazione Landau-Lifshitz-Gilbert (LLG), ovvero un modello matematico di base che descrive come il vettore di magnetizzazione di un materiale ferromagnetico varia nel tempo sotto l'effetto di un campo magnetico esterno ed interno. In formula:

[14] L'iMTJ sta per *in-plane Magnetic Tunnel Junction*, è uno dei modelli principali di MTJ, dove si ha una magnetizzazione orizzontale rispetto alla superficie.
[15] Gli elementi SPICE puri sono i componenti circuitali di base che il simulatore SPICE mette a disposizione senza ricorrere a modelli comportamentali o linguaggi avanzati.

$$\frac{\mathrm{dM}}{\mathrm{dt}} = -\gamma M \times \mathrm{H_{eff}} + \frac{\alpha}{M_s} M \times \frac{\mathrm{dM}}{\mathrm{dt}}$$

Dove:

-M = Vettore di magnetizzazione

-γ = Rapporto giromagnetico (costante che lega momento magnetico e momento angolare)

-H_{eff} = Campo magnetico efficace (somma di campi esterni, anisotropici, di scambio)

- α = Costante di smorzamento di Gilbert (determina quanto velocemente il sistema torna allo stato di equilibrio)

- M_s = Magnetizzazione di saturazione

2) Modello di Vatankhahghadim, Huda e Sheikholeslami: utilizza sempre l'equazione LLG ma aggiunge il termine di coppia di Slonczewski, causato dal disturbo introdotto dalla corrente polarizzata in spin. Questa nuova equazione (LLGS) richiede un periodo di simulazione più lungo, ma prevede il comportamento transitorio con un grado di accuratezza maggiore.

Tuttavia, entrambi i modelli hanno riconosciuto 4 zone differenti per le operazioni di commutazione, ciò è stato fatto per facilitare l'implementazione della modellazione dinamica. La dinamica della magnetizzazione è non lineare e complessa conviene quindi suddividere lo switching in regioni operative distinte, ognuna di esse ha condizioni fisiche differenti, perciò, possono essere trattate con set di equazioni più semplici e convergenti.

Per il primo modello che utilizza la LLG, le regioni sono basate sul valore dell'angolo magnetico (angolo tra il vettore di magnetizzazione e l'asse facile) rispetto all'angolo critico. Le regioni sono definite:

$\theta = 0$ → Allineamento perfettamente parallelo

$\theta > \theta_{th}$ → Angolo supera la soglia, inizio instabilità

$\theta < \theta_c$ → Ancora nello stato stabile vicino al minimo

$\theta > \theta_c$ → Regione di switch verso l'altro stato

Per il secondo modello che utilizza la LLGS, le regioni dipendono dal valore della corrente di scrittura I rispetto alla corrente critica I_c.

Le regioni sono definite:

$|I| > I_c^+$ →Corrente maggiore della soglia positiva

$|I| < I_c^+$ →Corrente minore della soglia positiva

$|I| > I_c$ →Corrente sopra la soglia generale

$|I| < I_c$ →Corrente sotto la soglia

Dividere il comportamento in 4 regioni rende la simulazione più robusta, più chiara e consente di modellare lo switch gradualmente, senza saltare bruscamente da "0" a "1". [11]

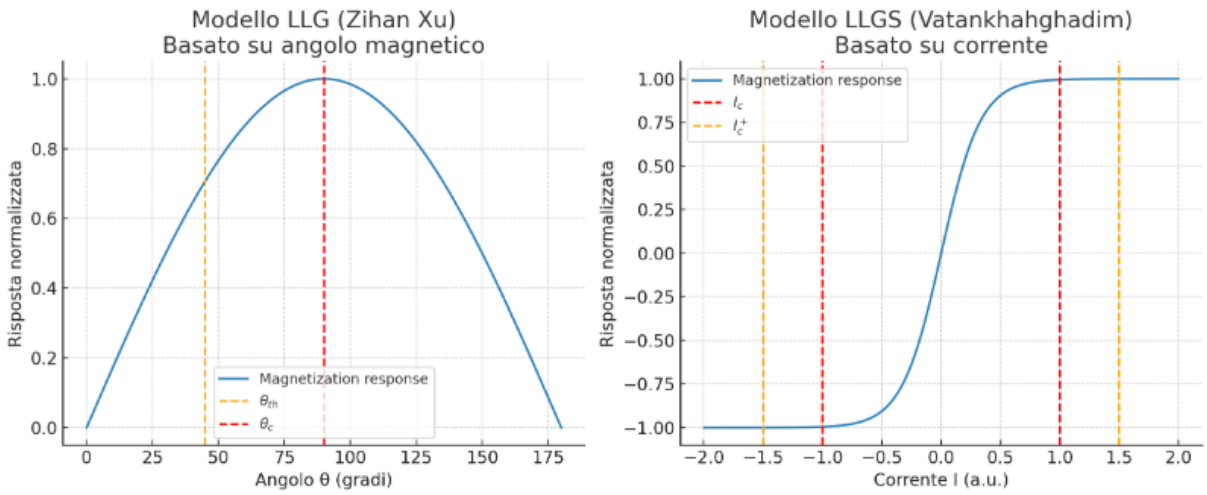


Figura 3.12: A sinistra LLG, a destra LLGS

Conclusioni e Risultati

Nelle conclusioni gli autori sottolineano che entrambi i modelli dinamici considerati sono stati simulati con successo, mostrando il comportamento atteso e confermando la validità dell'approccio. Le memorie STT-MRAM, pur offrendo numerosi vantaggi come non volatilità, velocità e resistenza alle radiazioni, presentano ancora alcune criticità da risolvere. In particolare, la latenza e il consumo energetico elevati durante la scrittura rappresentano un ostacolo importante, poiché il processo di commutazione è di natura stocastica e richiede ampi margini temporali per garantire il corretto funzionamento.

Per affrontare queste problematiche, la ricerca si sta orientando verso più direzioni: lo sviluppo di circuiti di scrittura a basso consumo, l'introduzione di tecniche di sensing più

accurate per ridurre il tasso di errore in lettura e l'adozione di circuiti periferici in grado di rendere la memoria più resistente alle radiazioni.

Allo stesso tempo, vengono studiate modifiche alla struttura stessa degli MTJ, come l'aumento del numero di strati, con l'obiettivo di eliminare la necessità di una cella di riferimento e ottenere operazioni di lettura prive di disturbi.

Pur essendo una tecnologia promettente, la STT-MRAM richiede ancora ottimizzazioni sia a livello di dispositivo che di circuito per raggiungere appieno le prestazioni necessarie nelle applicazioni future.

Capitolo 4

Applicazioni e integrazione delle NVM emergenti

4.1 Utilizzo nei dispositivi embedded e Iot

Per anni le memorie emergenti (MRAM, ReRAM, FeRAM, PCM) sono state viste come tecnologie rivoluzionarie in grado di unire la persistenza della Flash con la velocità e la resistenza della DRAM. Tuttavia, nonostante i vantaggi tecnici, il loro successo commerciale è stato limitato. Un esempio emblematico è Intel Optane (PCM) che, pur offrendo ottime prestazioni non è riuscito a diffondersi per motivi economici. La produzione su piccola scala mantiene i costi elevati, riduce la domanda e impedisce il raggiungimento di volumi sufficienti a sostenere il mercato. Di fronte a queste difficoltà, l'attenzione si è spostata verso le applicazioni embedded, dove le memorie tradizionali (come embedded flash) non riescono più a scalare sotto i 28nm. Le NVM emergenti offrono invece tempi di accesso rapidi, maggiore resistenza, efficienza energetica e migliore scalabilità, rendendole ideali per microcontrollori e applicazioni in settori come automotive, IoT e Edge AI.

La transizione verso l'integrazione embedded è già in corso, e le applicazioni trainanti richiedono caratteristiche specifiche:

Automotive → In questo settore le unità di controllo elettroniche e i sistemi avanzati di guida assistita richiedono una memoria con tolleranza alle alte temperature, alta resistenza e che garantisca cicli di scrittura ripetuti.

IoT → Nel settore IoT i dispositivi embedded richiedono bassissimo consumo e ritenzione dati prolungata.

Edge AI → Nell'Edge AI la latenza diventa una metrica prestazionale critica. Una memoria embedded più veloce può migliorare significativamente le prestazioni di inferenza e consentire il processo decisionale in tempo reale direttamente sul dispositivo.

Anche se non ancora pronte a sostituire DRAM e NAND su larga scala, le memorie emergenti stanno dimostrando la loro viabilità commerciale nel settore embedded. [12]

4.1.1 STT-MRAM Deep Learning Model for IoT Applications

L'apprendimento automatico è una delle tecnologie chiave per migliorare l'efficacia delle applicazioni per l'*Internet of Things*[16] (IoT). L'articolo che verrà analizzato propone un modello *hardware-aided* [17]basato su STT-MRAM per accelerare il deep learning in applicazioni IoT, in particolare nella *vision recognition*[18]. L'IoT fisicamente connesso alla STT-MRAM rappresenta una nuova tecnologia e un'ottima intersezione per lo sviluppo di tecniche di comunicazione e microelettromeccaniche.

Il problema di fondo è che i metodi software per eseguire CNN su dispositivi IoT sono limitati dalle risorse hardware, soprattutto dalla memoria. Le memorie tradizionali (come SRAM o SDRAM) consumano molta energia e non garantiscono sempre la velocità necessaria. La STT-MRAM, invece, offre vantaggi cruciali: basso consumo, non volatilità, alta resistenza e integrazione su nodi tecnologici avanzati, caratteristiche ideali per i dispositivi embedded e IoT.

Per fornire un riconoscimento visivo adeguato, la CNN esegue frequentemente operazioni di convoluzione sull'immagine di input. Tuttavia, le operazioni di convoluzione frequenti causerebbero un elevato consumo energetico e influirebbero sul tempo di riconoscimento. Di conseguenza un modello progettato a basso consumo energetico è altamente auspicabile per le implementazioni IoT.

[16] Con questo termine si indica l'insieme di oggetti fisici connessi a Internet e capaci di raccogliere, trasmettere ed elaborare dati
[17] Indica un approccio in cui alcune funzioni che normalmente verrebbero eseguite solo via software vengono invece affidate o accelerate da circuiti hardware dedicati
[18] Indica la capacità di un sistema di analizzare e identificare oggetti, persone, testi in immagini o video

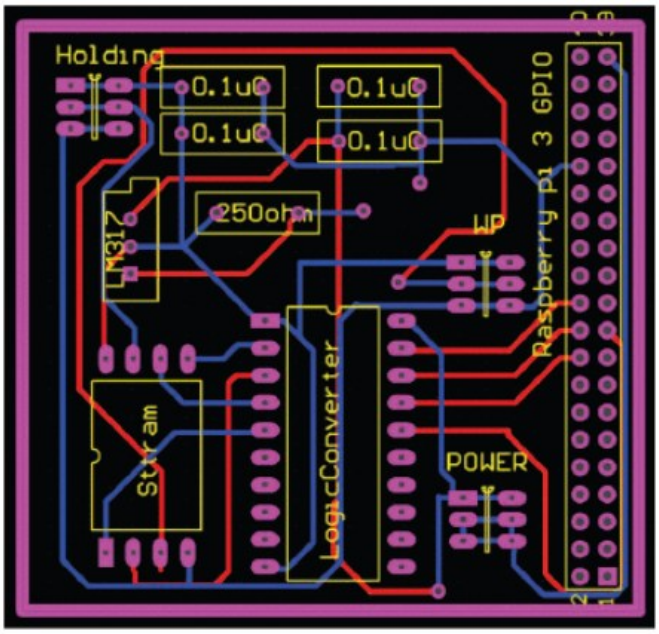


Figura 4.1: Circuito di controllo STT-MRAM

Progettazione

È stato progettato un circuito di controllo STT-MRAM vedi Fig.4.1, che viene accoppiato con una piattaforma hardware (FPGA[19]) tramite GPIO[20] (General-Purpose Input/Output) per simulare un segnale SPI[21], vedi Fig.4.2.

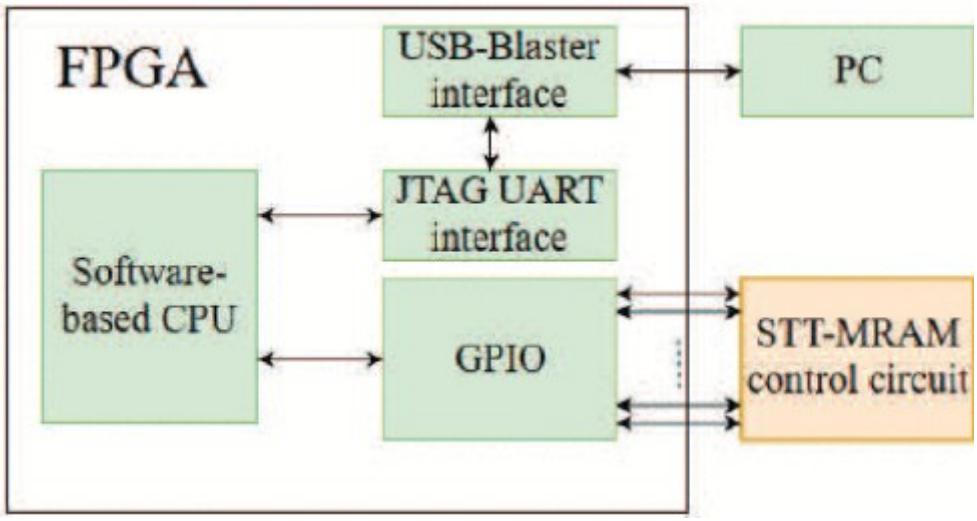


Figura 4.2: Schema del circuito di controllo STT-MRAM collegato a un FPGA

[19] È un dispositivo programmabile che permette di implementare circuiti digitali personalizzati direttamente nell'hardware, senza dover costruire un chip fisico nuovo
[20] È un pin presente su molti microcontrollori, programmabile per funzionare sia come ingresso che come uscita
[21] È un metodo per far scambiare dati digitali tra microcontrollori, FPGA, sensori e altri dispositivi periferici

Per eseguire il modello di apprendimento profondo CNN, abbiamo applicato NIOS II[22] nel FPGA come CPU basata su software, dove GPIO viene utilizzato dal processore NIOS II per comunicare con la STT-MRAM tramite il circuito di controllo vedi Fig. 4.3. Nel calcolo del riconoscimento visivo, è stato costruito un modello CNN multistrato per eseguire il deep learning e utilizzato il database MNIST come target per la valutazione.

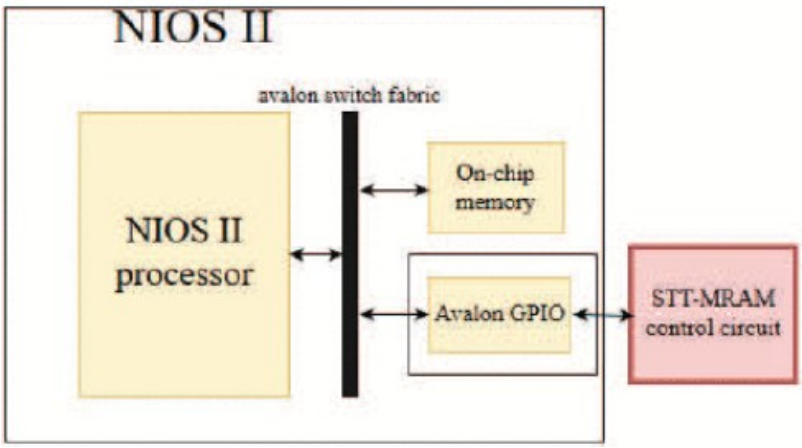


Figura 4.3: Schema del circuito di controllo STT-MRAM collegato a un FPGA

Infine, dato che i coefficienti o i pesi dei filtri del modello CNN devono essere memorizzati in modo valido e frequentemente accessibili, la STT-MRAM è stata selezionata come memoria non volatile per questo scopo. Grazie a questa progettazione, si può accelerare il tempo di esecuzione di almeno qualche secondo e migliorare l'accuratezza delle previsioni in unità di decine di migliaia di dati rispetto ai dispositivi IoT fisicamente collegati alla SDRAM.

L'integrazione tra STT-MRAM e FPGA rappresenta quindi una tendenza promettente per i dispositivi IoT di nuova generazione, grazie a prestazioni più elevate, consumi ridotti e memoria persistente. Con l'ingresso di STT-MRAM nella produzione di mazza, si prevede una diffusione sempre maggiore di architetture ibride di questo tipo. [13]

[22] È un processore soft-core programmabile sviluppato da Intel che viene implementato all'interno di FPGA

4.2 Integrazione nelle architetture di sistema (memorie artificiali, neuromorphic computing)

I dispositivi memristivi rappresentano una delle tecnologie più promettenti per emulare il funzionamento del cervello umano. Questi dispositivi, grazie alla loro capacità di modulare la resistenza elettrica e di mantenere in memoria gli stati raggiunti, possono replicare meccanismi fondamentali della biologia, come la plasticità sinaptica e l'attivazione neuronale. Ciò li rende particolarmente adatti a costruire reti neurali artificiali più efficienti sul piano energetico rispetto ai sistemi tradizionali basati su architetture von Neumann.

Un aspetto centrale riguarda l'uso di array di memristori, spesso organizzati in strutture a crossbar, che permettono di eseguire operazioni matriciali direttamente in memoria, abbattendo il cosiddetto memory wall della convenzionale architettura di von Neumann. Questo approccio consente di superare i limiti dovuti al continuo trasferimento di dati tra memoria e processore, riducendo consumi energetici e latenza. Grazie a queste proprietà, i memristori sono considerati una base concreta per lo sviluppo di reti neurali artificiali e sistemi neuromorfici sempre più complessi, come perceptron multistrato, reti convoluzionali e reti spiking già accennate in precedenza.

I memristori sono classificati secondo due criteri principali: a 2 terminali (più semplici) o a 3 terminali (con gate aggiuntivo, che offre maggiore controllo), e secondo il comportamento in termini di memoria ovvero:

-Non volatili, che mantengono lo stato anche senza alimentazione o Volatili, che lo perdono. La distinzione si estende anche su base digitale (solo due statu resistivi, SET/RESET definiti) o analogica (variazioni multipole e progressive di resistenza).

Questo spettro di comportamento rende i memristori estremamente flessibili e potenzialmente efficaci come elementi di storage e computazione nei sistemi neuromorfici.

Dal livello dispositivo si passa a sistemi più complessi: i chip neuromorfici veri e propri, che integrano array memristivi con circuiti periferici ottimizzati. Sebbene i memristori offrano vantaggi promettenti, è solo attraverso un'ottimizzazione coordinata tra materiali, architettura del circuito e algoritmi di apprendimento che si possono ottenere prestazioni realmente competitive.

Rimangono tuttavia sfide significative: i memristori soffrono di variabilità intrinseca, rumore, instabilità nei cicli di scrittura e mancanza di standardizzazione produttiva per impiego industriale su larga scala. Superare questi limiti richiede soluzioni che abbraccino la scala dei

materiali, il design architetturale e la robustezza algoritmica, con lo scopo di realizzare sistemi neuromorfici affidabili, scalabili ed efficienti. In quest'ottica, i sistemi neuromorfici basati su memristori, nonostante le difficoltà, possano giocare un ruolo cruciale nel futuro dell'edge computing, dell'AI embedded e dei dispositivi intelligenti ultra-efficienti, vedi Fig. 4.4. [14]

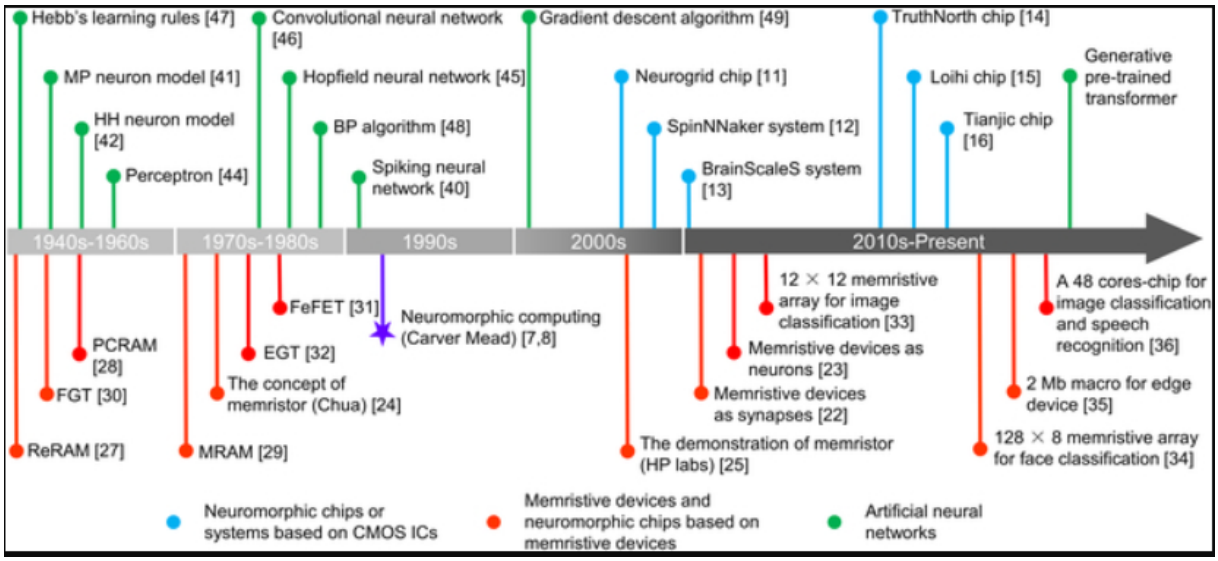


Figura 4.4: Le pietre miliari del calcolo neuromorfico

4.3 Impatto su datacenter, HPC e applicazioni AI

L'impatto delle memorie non volatili emergenti non si limita ai dispositivi embedded o al settore mobile, ma si estende anche a livello infrastrutturale, coinvolgendo datacenter, sistemi di calcolo ad alte prestazioni (HPC) e applicazioni di intelligenza artificiale. Nei data center, che rappresentano oggi una delle principali fonti di consumo energetico a livello globale, l'introduzione di tecnologie come PCM, ReRAM e STT-MRAM può ridurre significativamente i costi operativi e l'impronta carbonica. Questo è possibile grazie a una combinazione di caratteristiche uniche: consumi ridotti, elevata densità di integrazione, persistenza dei dati e maggiore affidabilità rispetto a DRAM e NAND Flash. Tali vantaggi consentono di ottimizzare lo storage, migliorare l'efficienza delle operazioni cloud e supportare la crescente richiesta di scalabilità per applicazioni di big data e servizi digitali avanzati.

Nel campo dell'HPC, l'adozione di memorie emergenti rappresenta una risposta concreta al *memory wall* nominato nel sottocapitolo precedente, ovvero il divario crescente tra la velocità di calcolo dei processori e la capacità della memoria di fornire dati con la stessa rapidità ed efficienza energetica. Le nuove memorie possono contribuire a ridurre i costi energetici, che

nei supercomputer di classe exascale derivano in larga parte dalla gestione della memoria, e consentire architetture ibride in cui DRAM e NVM coesistono per garantire al tempo stesso velocità, persistenza e capacità di storage. Inoltre, la possibilità di realizzare modelli di *in-memory* computing riduce la necessità di spostare continuamente i dati tra memoria e processore, con un conseguente guadagno in prestazioni e in efficienza.

Infine, nelle applicazioni di intelligenza artificiale, le memorie emergenti offrono benefici che vanno ben oltre l'efficienza energetica. L'addestramento e l'esecuzione di reti neurali profonde richiedono enormi quantità di dati e frequenti accessi in lettura e scrittura: in questo scenario, la persistenza, la velocità e l'elevata resistenza delle NVM permettono di ridurre la latenza, supportare modelli sempre più complessi e abilitare nuove forme di accelerazione hardware.

4.3.1 Architettura di acceleratore AI ottimizzato con STT-MRAM

Un acceleratore AI è un hardware specializzato progettato per eseguire in maniera molto più rapida ed efficiente i calcoli richiesti dagli algoritmi di intelligenza artificiale, in particolare dal deep learning.

Nel 2021 è stata presentata un'innovativa architettura di acceleratore AI ottimizzato con STT-MRAM su chip, con l'obiettivo di ottenere prestazioni elevate riducendo al contempo spazio e consumo energetico. A differenza del classico approccio basato su SRAM, qui la memoria viene progettata su misura per le esigenze specifiche del deep learning.

Il cuore della proposta è un sistema di *buffer on-chip*[23] che utilizza la STT-MRAM come memoria *scratchpad*[24], affiancata a un *core* riconfigurabile[25].

Per calibrare al meglio il comportamento della memoria, è stata sviluppata una metodologia guidata da un modello analitico, basato sulla durata prevista dei pesi dei modelli AI e delle mappe di attivazione nella memoria. Da questo modello ne consegue la possibilità di regolare la stabilità termica (*thermal stability factor*, Δ) della memoria.

[23] Memoria "vicina" al processore, integrata direttamente nel chip. Serve come zona di appoggio temporanea per i dati che devono essere usati subito, senza dover andare a recuperarli da memorie più lente.
[24] È un tipo di memoria molto veloce e gestita direttamente dal software/hardware, usata per contenere dati temporanei necessari a calcoli immediati.
[25] La parte di calcolo non è fissa, ma può essere adattata a seconda del tipo di rete neurale o dell'applicazione AI.

Un elevato Δ garantisce lunga conservazione degli stati, ma non servendo in un acceleratore AI dove i dati variano in millisecondi o secondi, esso può essere ridotto opportunamente (tramite modifiche geometriche nel materiale dell'MTJ), riuscendo così ad aumentare la densità della memoria, riducendo i consumi e diminuendo i tempi di lettura e scrittura, non compromettendo l'accuratezza richiesta dalle reti neurali.

Un problema che viene affrontato è quello delle variazioni di processo e temperatura: il valore di Δ può oscillare tra chip e condizioni operative. Per garantire affidabilità, integrano un circuito intelligente chiamato *write-driver* adattivo che, in base a un monitoraggio continuo, regola dinamicamente la corrente di scrittura nei casi peggiori.

L'architettura ottenuta, chiamata STT-AI, è implementata con tecnologia a 14nm. Rispetto ad un acceleratore tradizionale basato su SRAM e alle stesse prestazioni di precisione, l'approccio STT-AI raggiunge una riduzione del 75% dell'area fisica e un risparmio energetico del 3%, vedi Fig. 4.5. [15]

Module	Details	Area (mm^2)	Dynamic Power (mW)	Leakage Power (mW)
Functional Core	Reconfigurable core with 42x42 MACs	4.08	954	0.91
SRAM Block	12 MB SRAM global memory	16.2	48.98	0.21
STT-MARM (Δ=27.5)	12 MB MRAM global memory	1.01	17.61	0.08
STT-MRAM (Δ=17.5, Δ=27.5)	6 MB MRAM (Δ=17.5) 6 MB MRAM (Δ=27.5)	0.93	13.75	0.06
SRAM ScratchPad (for MRAM)	52 KB (two 26KB blocks with CLK/power gating)	0.069	0.2	8E-4
Baseline Accelerator (SRAM-based)	Functional Core and SRAM (Row 3 above)	20.28	1003	1.13
STT-AI Accelerator	Functional core and STT-RAM (Row 4 above)	5.09	972	0.99
STT-AI Ultra Accelerator	Functional Core and STT-RAM (Row 5 above)	5.0	968	0.98

Figura 4.5: Dettagli dell'acceleratore a 14nm

Da ciò si evince come una NVM emergente possa essere personalizzata per usi specifici, affiancando matematica, architettura e circuitazione, come si possa ottenere una riduzione massiccia di area e un miglioramento energetico con mantenimento della funzionalità AI.

Capitolo 5

Analisi comparativa e prospettive future

5.1 Confronto tra le principali tecnologie NVM emergenti

Dopo aver analizzato nel dettaglio i principi di funzionamento e le applicazioni specifiche delle singole tecnologie nei capitoli precedenti, è essenziale procedere a un confronto diretto per comprendere il loro posizionamento all'interno della gerarchia di memoria futura.

L'analisi si concentra sulle tre tecnologie principali trattate: STT-MRAM, PCM e OxRAM. Nonostante tutte appartengano alla categoria delle memorie non volatili emergenti, le loro caratteristiche fisiche e prestazionali le rendono adatte a scopi differenti, piuttosto che concorrenti diretti per la stessa nicchia di mercato, vedi Fig. 5.1.

5.1.1 Analisi dei parametri prestazionali

- **Densità e Scalabilità**: Come analizzato nel Cap. 3, la STT-MRAM si basa sulla giunzione magnetica a tunnel (MTJ), una struttura multistrato complessa. Sebbene offra prestazioni eccellenti, la complessità geometrica della MTJ rende la scalabilità a nodi tecnologici piccolissimi più ardua rispetto alle controparti resistive. Al contrario, le celle RRAM e PCM, basandosi rispettivamente su semplici strutture MIM e sul riscaldamento di un volume di materiale calcogenuro, mostrano un potenziale di densità superiore, permettendo una maggiore capacità di archiviazione per unità di area.
- **Velocità e Latenza**: La STT-MRAM eccelle in velocità, posizionandosi come la candidata principale per sostituire le memorie di lavoro (SRAM e DRAM). La sua commutazione basata sullo spin degli elettroni è estremamente rapida e non richiede i cambiamenti strutturali significativi del materiale necessari nelle altre tipologie. La PCM, pur essendo veloce, è limitata fisicamente dai tempi necessari per la transizione di fase: l'operazione di reset richiede la fusione e il rapido raffreddamento del materiale, mentre il set richiede un tempo prolungato per la cristallizzazione.
- **Resistenza**: Un fattore critico evidenziato nel confronto è la resistenza ai cicli di scrittura. La STT-MRAM dimostra una resistenza eccezionale, nell'ordine dei 10^{15}

cicli, rendendola praticamente immune all'usura per la maggior parte delle applicazioni embedded. La PCM, invece, soffre dello stress termico ripetuto che può degradare il materiale calcogenuro nel tempo, offrendo un'endurance inferiore seppur nettamente migliore rispetto alle Flash tradizionali.

- **Efficienza Energetica**: l'efficienza della STT-MRAM è stata evidenziata nel caso studio dell'acceleratore AI trattato nel Cap. 4. L'implementazione di questa memoria ha permesso di ridurre l'area fisica del 75% e il consumo energetico del 3% rispetto a una soluzione basata su SRAM. Questo conferma la sua idoneità per applicazioni a basso consumo come l'IoT.

5.1.2 Verso una specializzazione delle tecnologie

Dall'analisi emerge chiaramente che non esiste una "vincitrice unica" in grado di sostituire universalmente tutte le memorie esistenti. Si delinea piuttosto uno scenario di specializzazione:

1) **STT-MRAM come Memoria Unificata Embedded**: Grazie alla sua velocità, resistenza e compatibilità con i processi CMOS, è la soluzione ideale per sostituire la SRAM nelle cache di livello superiore (L2, L3) e la Flash nei microcontrollori.
2) **PCM e RRAM per Storage Class Memory e Neuromorphic Computing**: La loro capacità di modulare la conduttanza in modo analogico e su più livelli, come visto negli esperimenti sulla plasticità sinaptica (STDP) nel Cap. 3, le rende le tecnologie d'elezione per il calcolo neuromorfico. Inoltre, la loro alta densità le posiziona perfettamente come Storage Class Memory, colmando il divario di prestazioni tra la memoria principale (DRAM) e l'archiviazione di massa (SSD).

Caratteristica	STT-MRAM	PCM	RRAM (OxRAM)
Meccanismo Fisico	Spintronica (MTJ)	Cambiamento di fase (Calore)	Commutazione resistiva (Filamento)
Densità (Scalabilità)	Media (Limitata dalla complessità MTJ)	Alta (Scalabilità 1T1R/crossbar)	Molto Alta (Struttura semplice MIM)
Velocità (Latenza)	Molto Alta (vicina alla SRAM)	Media (limitata da raffreddamento)	Alta (Commutazione rapida)
Resistenza (Endurance)	Altissima (~ 10^15 cicli)	Limitata (stress termico)	Buona (superiore Flash, < MRAM)
Consumo Energetico	Basso (-3% energia vs SRAM)	Medio (corrente per riscaldare)	Basso (bassa tensione set/reset)
Applicazione Ideale	Cache L2/L3, Embedded, IoT	Storage Class Memory (SCM), Neuromorphic	Neuromorphic, Storage Class Memory (SCM)

Figura 5.1: Confronto tra le principali NVM

5.2 Sfide tecnologiche e limiti attuali

Nonostante le eccellenti caratteristiche teoriche evidenziate nel confronto precedente, l'adozione su larga scala delle memorie NVM emergenti deve affrontare ostacoli significativi, sia di natura fisica che economica.

- **La sfida della variabilità nelle OxRAM**: Come emerso dall'analisi sperimentale nel Cap.3, uno dei limiti critici delle memorie resistive è l'elevata variabilità della conduttanza. Per le applicazioni di memoria standard, questa fluttuazione rappresenta un serio problema di affidabilità, poiché riduce il margine di rumore tra lo stato logico '0' e '1', richiedendo complessi circuiti di correzione degli errori. Tuttavia, è interessante notare come questo "difetto" possa trasformarsi in un punto di forza in ambiti specifici: come dimostrato nell'esperimento sulla rilevazione delle auto, la variabilità stocastica aiuta le reti neurali a esplorare uno spazio dei pesi più ampio, migliorando l'apprendimento.
- **La natura stocastica della STT-MRAM**: La sfida principale qua, risiede nella scrittura. I modelli dinamici basati su LLG e LLGS analizzati nel Cap.3, evidenziano che la commutazione di spin è un processo intrinsecamente stocastico e non deterministico. Ciò implica che per garantire un basso tasso di errore di scrittura, è necessario applicare correnti più elevate o impulsi più lunghi, penalizzando il consumo energetico e la latenza. Per mitigare questo problema, non compromettendo le prestazioni, è necessario, come visto nel caso dell'acceleratore AI, implementare dei circuiti periferici. Questi circuiti sono in grado di monitorare le variazioni di processo e temperatura in tempo reale, regolando dinamicamente la corrente di scrittura solo quando necessario.

5.3 Prospettive evolutive e roadmap tecnologiche

- **Storage Class Memory (SCM)**: La roadmap tecnologica posiziona le NVM (in particolare PCM e RRAM) come la soluzione definitiva per creare il livello di SCM. Attualmente esiste un enorme divario prestazionale (in termini di latenza) tra la memoria principale (DRAM, veloce ma volatile) e l'archiviazione secondaria (SSD, non volatile ma lenta). Le NVM si inseriscono esattamente in questo spazio, offrendo una persistenza dei dati con velocità vicine alla DRAM. Questo permetterà ai futuri

sistemi operativi di trattare l'archiviazione permanente quasi come se fosse memoria RAM, rivoluzionando la gestione dei database e dei file system.

- **Oltre Von Neumann**: La prospettiva più rivoluzionaria riguarda il superamento del "Memory Wall" dell'architettura di Von Neumann, discusso nel Cap. 4. Invece di spostare continuamente i dati tra memoria e processore, operazione molto costosa in termini di energia e tempo, il futuro risiede nell'In-Memory Computing, ovvero eseguire calcoli direttamente all'interno degli array di memoria. In questo scenario, le OxRAM e le PCM non fungono solo da magazzino dati, ma agiscono come sinapsi artificiali. La loro capacità di modulare la resistenza in modo analogico (e non solo binario), è la chiave per realizzare hardware neuromorfico in grado di eseguire algoritmi di Intelligenza Artificiale con un'efficienza energetica irraggiungibile per le architetture tradizionali.

Capitolo 6

Conclusioni

6.1 Sintesi dei principali risultati

Il percorso di analisi svolto in questa tesi ha permesso di tracciare un quadro chiaro dell'evoluzione delle memorie non volatili. Si è partiti dalla presa d'atto di un limite strutturale: le tecnologie tradizionali, ovvero DRAM e Flash, hanno raggiunto un collo di bottiglia fisico (scaling) e prestazionale (memory wall) che non è più sostenibile per le moderne esigenze di calcolo.

Dall'analisi fisica delle tecnologie emergenti condotta nel Cap.3, è emerso un risultato fondamentale: la diversificazione. A differenza del passato, dove un'unica tecnologia (come la Flash) dominava l'intero mercato non volatile, le nuove NVM si differenziano per meccanismi fisici: spintronica per la STT-MRAM, filamenti conduttivi per la RRAM e transizione di fase per la PCM, che ne dettano impieghi specifici e non sovrapponibili.

Il risultato più significativo, tuttavia, riguarda l'integrazione applicativa discussa nel Cap.4. Lo studio dei casi d'uso ha dimostrato che le presunte "imperfezioni" di queste memorie posso trasformarsi in vantaggi strategici.

In particolare:

- Nell'ambito delle reti neurali *Spiking*, si è visto come la variabilità stocastica della OxRAM, solitamente considerata un difetto in ambito digitale, favorisca in realtà l'apprendimento e la plasticità sinaptica
- Nell'ambito embedded e IoT, l'integrazione della STT-MRAM in acceleratori hardware (come dimostrato nell'architettura STT-AI) ha validato la possibilità di ridurre drasticamente l'area del chip e il consumo energetico, superando i vincoli delle implementazioni basate su SRAM.

In sintesi, il lavoro conferma che le NVM emergenti hanno superato la fase di pura ricerca sui materiali per entrare in quella validazione architetturale, dimostrandosi pronte per risolvere problemi specifici che le memorie classiche non possono più affrontare.

6.2 Considerazioni sull'impatto della nuova tecnologia

Concludendo, è possibile estendere lo sguardo oltre i dati tecnici per valutare l'impatto macroscopico che queste tecnologie avranno sull'ecosistema digitale. L'adozione delle NVM emergenti non rappresenta un semplice aggiornamento incrementale, ma un vero e proprio cambio di paradigma che abilita tre trasformazioni cruciali.

La prima è la diffusione capillare dell'**intelligenza artificiale**. Portare l'inferenza e l'apprendimento direttamente sui dispositivi finali, grazie a memorie come la STT-MRAM che offrono persistenza e velocità a baso consumo, svincola l'IoT dalla dipendenza costante dal cloud. Questo ha implicazioni dirette sulla privacy dei dati e sulla reattività dei sistemi critici, come nel settore automotive.

La seconda considerazione riguarda la **sostenibilità energetica**. In un'epoca in cui i Data Center consumano una quota rilevante dell'energia globale, la transizione verso memorie che non richiedono alimentazione per mantenere i dati e che permettono il calcolo in-memory riduce drasticamente l'impronta carbonica delle infrastrutture digitali.

Infine, assistiamo alla fine della rigida separazione tra **memoria ed archiviazione**. La maturazione di tecnologie come la PCM sta sfumando il confine tra la memoria di lavoro e lo storage. Questo porterà a una ridefinizione delle architetture dei calcolatori, semplificando la gerarchia di memoria che abbiamo studiato nei primi capitoli e permettendo lo sviluppo di sistemi sempre più performanti.

In definitiva, il prossimo decennio sarà probabilmente definito dalla capacità delle nuove memorie di abbattere il muro che separa il dato dalla sua elaborazione.

6.3 Spunti per future richerche

Il lavoro svolto in questa tesi ha evidenziato la potenzialità delle memorie non volatili emergenti, ma ha anche aperto la strada a nuovi interrogativi che meritano ulteriori approfondimenti sperimentali. Di seguito vengono proposti due filoni di ricerca che potrebbero estendere i risultati qui presentati.

1) **Ottimizzazione dei circuiti di scrittura per le STT-MRAM**: Come discusso nel Cap. 4, la natura stocastica della commutazione nella STT-MRAM impone margini di sicurezza ampi che penalizzano la latenza e il consumo energetico. Ricerche future potrebbero concentrarsi sulla progettazione di circuiti intelligenti ancora più avanzati rispetto a quelli analizzati. L'obiettivo sarebbe quello di monitorare in tempo reale l'avvenuta commutazione della cella MTJ per interrompere la corrente di scrittura nell'istante esatto in cui il bit viene capovolto, riducendo drasticamente lo spreco energetico e migliorando l'affidabilità del dispositivo.
2) **Architettura di memoria ibrida gerarchica**: Infine, sebbene questa tesi abbia spesso confrontato le tecnologie singolarmente, il futuro potrebbe risiedere nella loro combinazione. Uno spunto interessante per il proseguimento dello studio riguarda la simulazione di architetture ibride su chip, dove una cache L1/L2 in SRAM (per la massima velocità) coopera con una cache L3 in STT-MRAM (per la densità) e una memoria principale PCM. Analizzare come i compilatori e i sistemi operativi possano gestire dinamicamente il posizionamento dei dati tra diversi livelli di memoria non volatile rappresenta una sfida aperta e di grande interesse per l'evoluzione dei sistemi embedded ad alte prestazioni.

"La strada è dunque tracciata: l'evoluzione delle memorie non riguarda più solo la capacità di archiviare il passato, ma la velocità con cui saremo in grado di elaborare il futuro."

Bibliografia

[1] Hennessy, J. L., & Patterson, D. A. (2019). *Computer Architecture: A Quantitative Approach* (6th ed.), Capitolo 5, sezione 5.1 (Memory Hierarchy).

[2] Hennessy, J. L., & Patterson, D. A. (2019). *Computer Architecture: A Quantitative Approach* (6th ed.), Capitolo 2 (Memory Hierarchy Design).

[3] Computer Systems: A Programmer's Perspective" – Randal E. Bryant, David R. O'Hallaron (3rd ed.), Capitolo 6 (The Memory Hierarchy) e sezione 6.6.1 (The Memory Mountain).

[4] Hennessy, J. L., & Patterson, D. A. (2019). *Computer Architecture: A Quantitative Approach* (6th ed.), Capitolo 1, sezione 1.2-1.4, pp. 18-26.

[5] Hennessy, J. L., & Patterson, D. A. (2019). *Computer Architecture: A Quantitative Approach* (6th ed.), Capitolo 2, sezione 2.2, pp. 78-92.

[6] Hennessy, J. L., & Patterson, D. A. (2019). *Computer Architecture: A Quantitative Approach* (6th ed.), Capitolo 2, sezione 2.2, pp. 92.

[7] N. Aswathy and N. M. Sivamangai, "Future Nonvolatile Memory Technologies: Challenges and Applications," *2021 2nd International Conference on Advances in Computing, Communication, Embedded and Secure Systems (ACCESS)*, Ernakulam, India, 2021, pp. 308-312.

[8] S. Raman Sundara Raman, 'A Review on Non-Volatile and Volatile Emerging Memory Technologies', Computer Memory and Data Storage. IntechOpen, Jan. 10, 2024.

[9] Zahoor, F., Azni Zulkifli, T.Z. & Khanday, F.A. Resistive Random Access Memory (RRAM): an Overview of Materials, Switching Mechanism, Performance, Multilevel Cell (mlc) Storage, Modeling, and Applications. *Nanoscale Res Lett* **15**, 90 (2020).

[10] E. Vianello *et al*., "Metal Oxide Resistive Memory (OxRAM) and Phase Change Memory (PCM) as Artificial Synapses in Spiking Neural Networks," *2018 25th IEEE International Conference on Electronics, Circuits and Systems (ICECS)*, Bordeaux, France, 2018, pp. 561-564.

[11] B. R. Mathew and N. M. Siva Mangai, "Analysis of Dynamic Models of a STT-MRAM," *2018 4th International Conference on Devices, Circuits and Systems (ICDCS)*, Coimbatore, India, 2018, pp. 259-262.

[12] Ethan Phillips, *May 29, 2025* “Emerging Memory Takes the Embedded Route”

[13] H. -J. Lai e Y. -T. Tsou, "Modello di apprendimento profondo di co-progettazione STT-MRAM per applicazioni IoT", *8a conferenza globale IEEE sull'elettronica di consumo (GCCE) del 2019* , Osaka, Giappone, 2019, pp. 421-422

[14] Yike Xiao, Cheng Gao, Juncheng Jin, Weiling Sun, Bowen Wang, Yukun Bao, Chen Liu, Wei Huang, Hui Zeng, Yefeng Yu. Recent Progress in Neuromorphic Computing from Memristive Devices to Neuromorphic Chips. 2024.

[15] arXiv:2104.02199 , “Designing Efficient and High-performance AI Accelerators with Customized STT-MRAM” , Kaniz Mishty, Mehdi Sadi, *6 Apr 2021.*