

# Introduzione alle memorie

Matteo Sonza Reorda

Politecnico di Torino  
Dip. di Automatica e Informatica



# Introduzione

**Ogni sistema di elaborazione contiene dispositivi per la memorizzazione di dati ed istruzioni.**

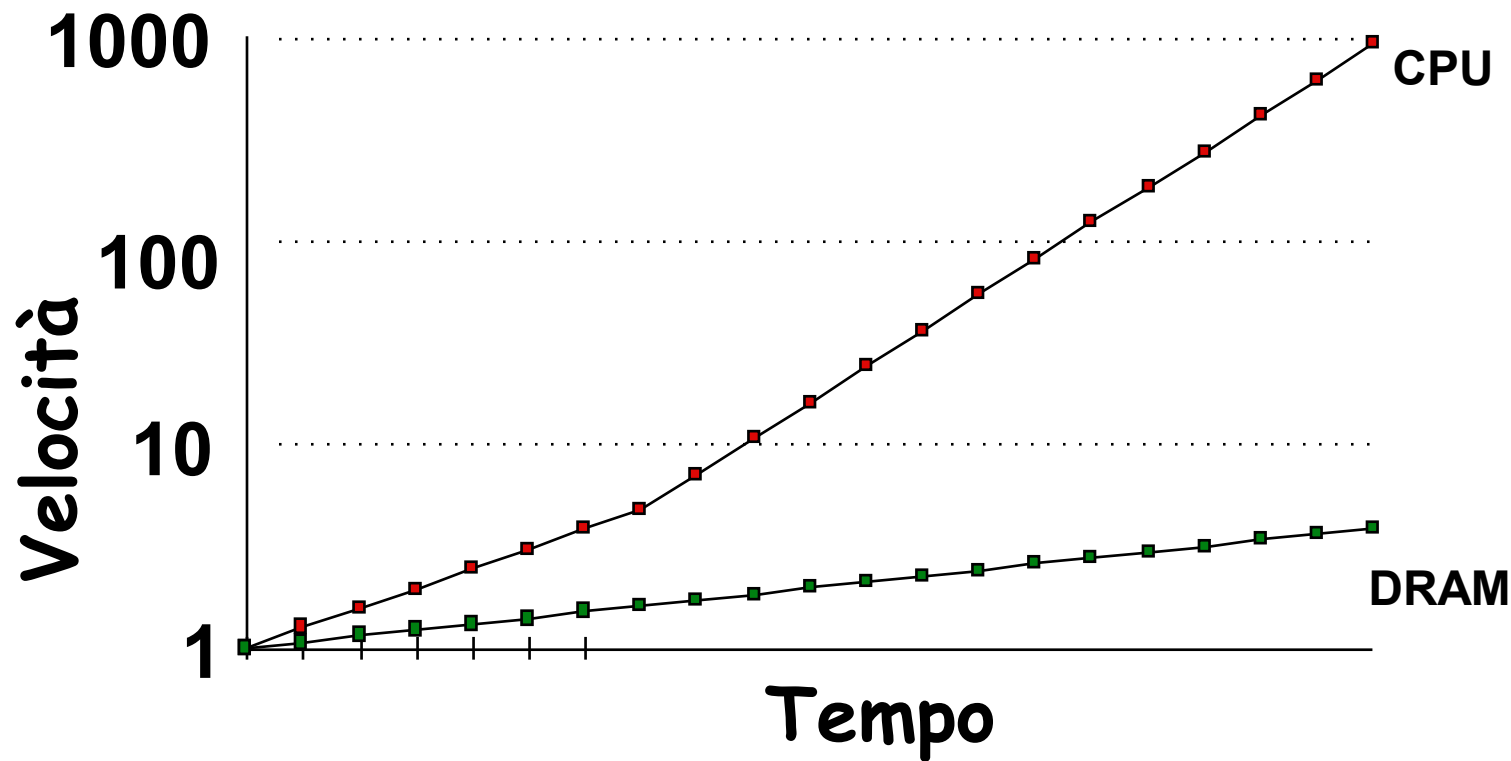
**L'insieme di tali dispositivi, e degli algoritmi per la loro gestione, costituisce il *sotto-sistema di memoria*.**

**Tale sotto-sistema deve essere realizzato in modo che**

- **il processore debba attendere il meno possibile per accedere a dati o istruzioni**
- **il costo del sistema di memoria sia minimo (il costo per bit delle memorie è proporzionale alla loro velocità).**

**Si deve quindi cercare un compromesso tra il costo del sotto-sistema di memoria e le sue prestazioni.**

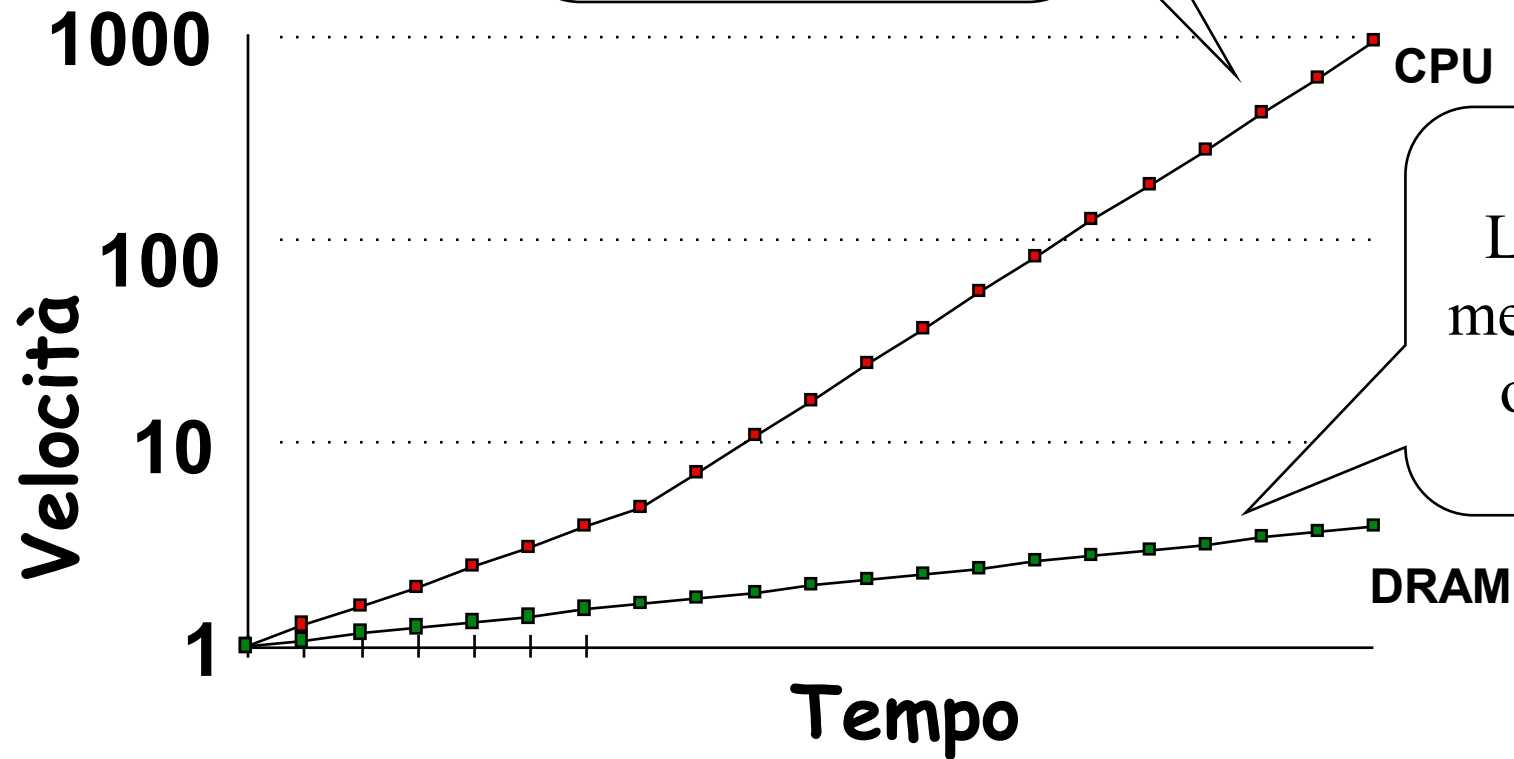
# Velocità del processore/velocità della memoria



# Velocità

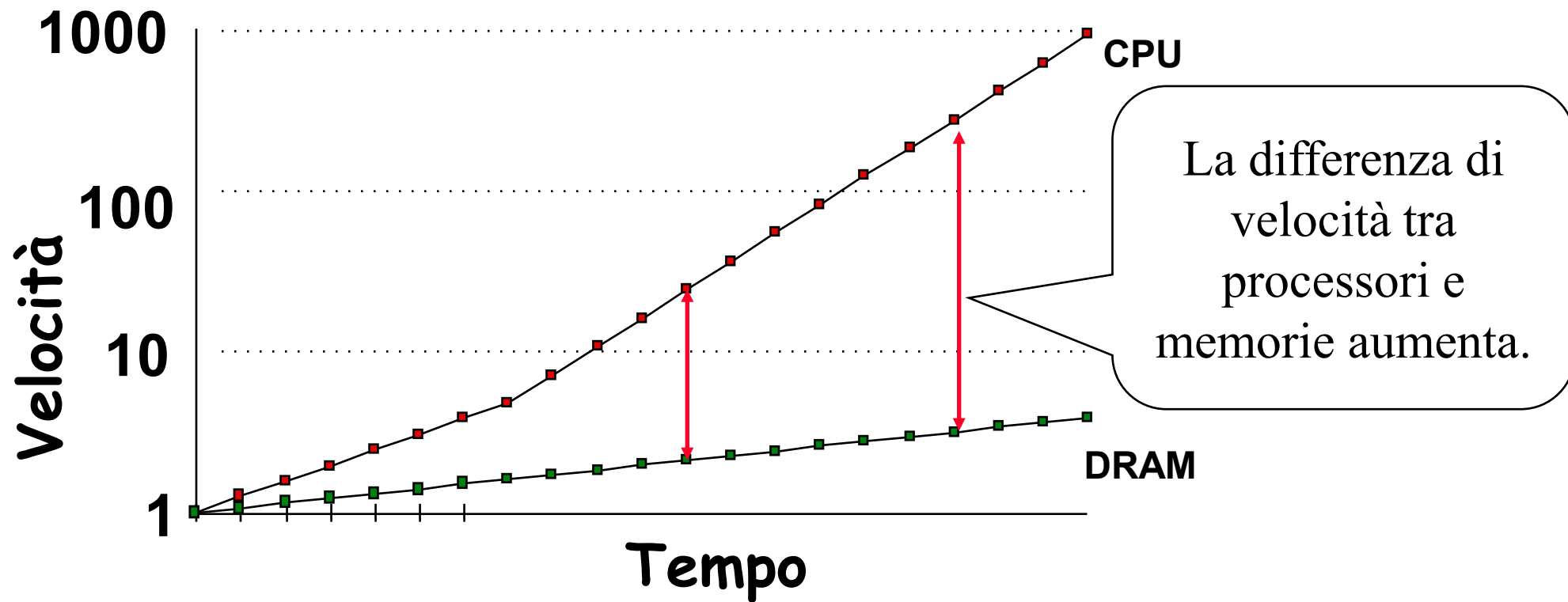
La velocità dei  
processori  
raddoppia ogni  
anno e mezzo.

# Processore/velocità memoria



La velocità delle  
memorie raddoppia  
ogni dieci anni.

# Velocità del processore/velocità della memoria



# Evoluzione delle memorie

L'evoluzione della tecnologia presenta 2 tendenze:

- riduzione del *costo* per bit
- riduzione (meno marcata) del *tempo di accesso*.

# **Livelli di memorie**

**La memoria di un calcolatore è normalmente organizzata in *livelli*.**

**Il numero di livelli dipende dal tipo di sistema e di applicazione.**

**Ogni livello è caratterizzato da un diverso tipo di memoria (in termini di velocità e dimensione, e quindi costo).**

# Obiettivo

**Ottimizzare il Sistema di Memoria in modo tale da minimizzare il tempo medio di accesso per i tipici programmi che utilizziamo (*workload*):**

- **Le informazioni sono distribuite in modo dinamico tra i vari livelli a seconda della relativa frequenza di accesso**
- **In tal modo si raggiungono prestazioni comparabili con quelle della memoria più veloce, a prezzi comparabili con quelli della memoria più lenta.**



# **Principio di località dei riferimenti**

**Un riferimento in memoria tende ad essere ripetuto dopo poco tempo:**

- **Località temporale**

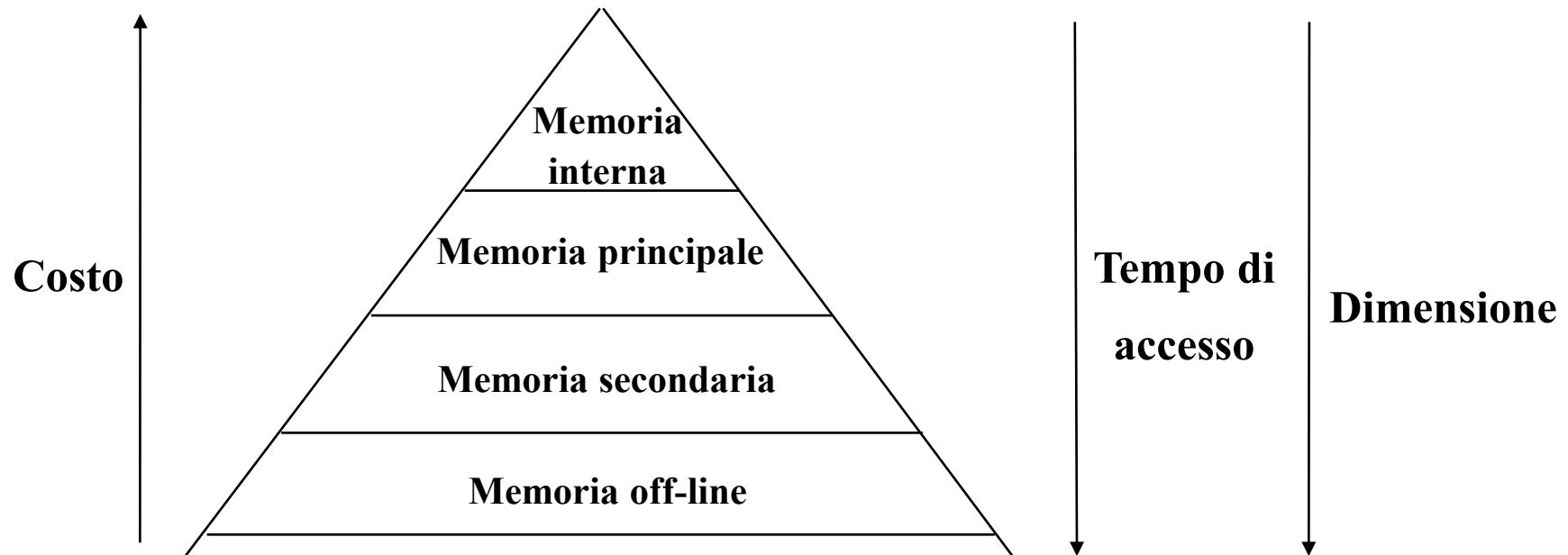
**Un riferimento in memoria tende ad essere a locazioni vicine a quelle usate recentemente**

- **Località spaziale.**

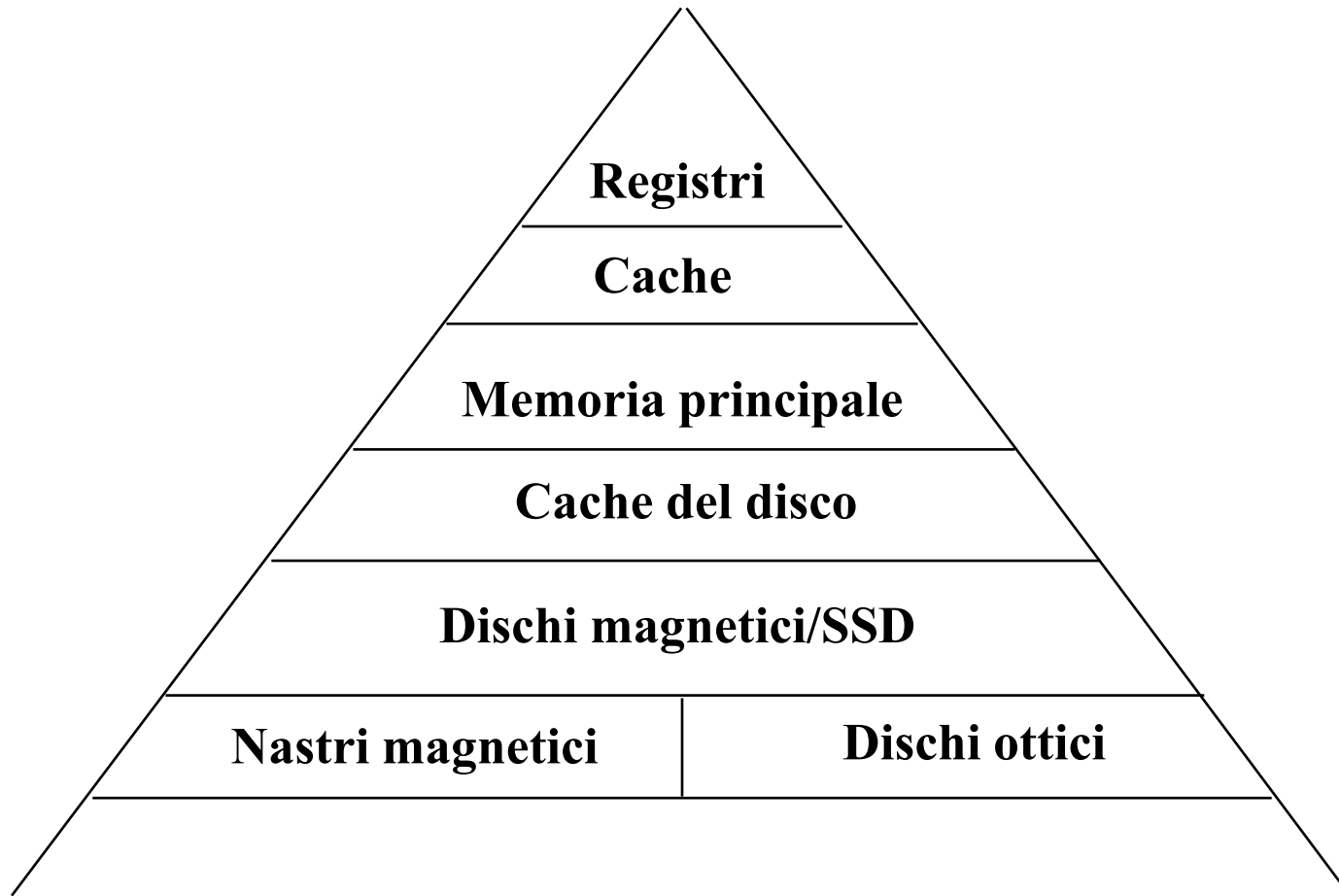
# **Implicazioni del principio di località**

- **Posso utilizzare piccole memorie veloci per mantenere i riferimenti in memoria più utilizzati dal processore**
- **Se i riferimenti utilizzati non possono essere soddisfatti nella memoria veloce creo un secondo livello di memoria un po' più grande e un po' meno veloce**
- **Posso reiterare questo ragionamento per N livelli di memoria:**
  - **Gerarchia di memoria**

# Gerarchia di memoria



# Situazione attuale



# Memoria interna alla CPU

**Corrisponde ai registri interni della CPU.**

**È caratterizzata da:**

- **alta velocità (comparabile con quella del processore): i tempi di accesso sono dell'ordine dei ns**
- **limitate dimensioni (al più qualche Kbyte).**

**È di solito realizzata tramite celle di *RAM statica*.**

# Memoria principale

**Ha dimensioni molto maggiori (da qualche Mbyte fino a qualche Gbyte) ma tempi di accesso più elevati, dell'ordine delle decine di ns.**

**È accessibile in modo diretto tramite indirizzi.**

**È normalmente realizzata sotto forma di circuito integrato.**

**È di solito realizzata tramite circuiti di *RAM dinamica*.**

# Memoria secondaria

**Ha dimensioni e tempi di accesso ancora maggiori, dell'ordine delle decine di ms.**

**Ha dimensioni che possono arrivare ai Tbyte.**

**Viene usata per memorizzare dati e programmi non immediatamente richiesti dal processore.**

**L'accesso è gestito da appositi programmi di interfaccia.**

**È normalmente realizzata sotto forma di *dischi magnetici* o memorie *Flash* (anche note come Solid State Disk, o SSD), o combinazioni dei due.**

**Le informazioni memorizzate nella memoria secondaria non vengono perse allo spegnimento del sistema.**

# Memoria Off-line

**Permette di memorizzare grandi moli di dato (Pbyte) con tempi di accesso elevati (decine di secondi).**

**In taluni casi l'accesso alla memoria off-line richiede l'intervento di un operatore.**

**Di solito è composta da *dischi ottici* o *nastri*.**



# Cache

**Le *cache* sono memorie estremamente veloci che si interpongono tra il processore e la memoria principale.**

**All'interno di una cache risiedono temporaneamente i dati/programmi utilizzati in quel momento.**

**Il loro uso è trasparente al programmatore e al Sistema Operativo.**

**Le cache permettono di aumentare la velocità di accesso ai dati nella memoria principale senza ricorrere per essa a memorie di tipo più costoso.**

# Strategia generale

**Nella progettazione di un sistema di memoria vengono tenuti in conto i seguenti punti:**

- **conviene che il sistema complessivo sia composto da memorie di tipo e costo diversi, rispondenti ai diversi usi che della memoria vengono fatti (*gerarchia di memoria*)**
- **la gestione della memoria deve essere il più possibile trasparente per il programmatore e l'utente (*memoria virtuale*)**
- **se il sistema è di tipo *multiprocessore*, ogni processore deve poter lavorare con la memoria al massimo della velocità e senza interferire con il lavoro degli altri.**

# **Classificazione delle tecnologie di memoria**

**È basata su vari parametri:**

- **costo**
- **velocità**
- **modi di accesso**
- **alterabilità**
- **durevolezza del contenuto**
- **affidabilità**
- **caratteristiche fisiche.**

# Costo

**Comprende anche il costo della circuiteria ed eventualmente del software per la gestione delle interfacce necessarie all'uso della memoria.**

**È normalmente misurato in dollari/bit o \$/Mbyte.**

# Modi di accesso

Sono principalmente:

- ***sequenziale***: le informazioni possono essere lette/scritte solo in un ordine prefissato; è il caso dei nastri
- ***diretto***: ogni blocco ha un indirizzo, e la ricerca nel blocco è sequenziale; è il caso dei dischi magnetici
- ***casuale***: ogni unità di dato ha un indirizzo, e le informazioni possono essere lette/scritte in qualsiasi ordine; il tempo di accesso è uguale per tutte le locazioni di memoria; è il caso delle memorie a semiconduttore
- ***associativo***: l'accesso avviene tramite un confronto tra il contenuto di ogni cella e quello specificato in una maschera; è il caso di taluni tipi di cache.

# Velocità

**Si esprime attraverso tre parametri**

- **il tempo di accesso**
- **il tempo di ciclo**
- **il tasso di trasferimento.**

# Tempo di accesso (o latenza)

**È il tempo che intercorre tra l'istante in cui all'unità di memoria giunge la richiesta di eseguire un'operazione (lettura o scrittura), e quello in cui tale operazione è eseguita.**

**È possibile che il tempo di accesso *in lettura* differisca da quello *in scrittura*.**

**Il tempo di accesso è di solito inversamente proporzionale al costo della memoria.**

# Tempo di ciclo

**È il tempo che deve intercorrere tra l'inizio di un ciclo di accesso alla memoria e l'inizio del ciclo successivo.**

**È chiaramente maggiore o uguale del tempo di accesso.**



# Tasso di trasferimento

È la velocità con la quale i dati possono esser trasferiti verso o dall'unità di memoria.

Per le memorie ad accesso casuale è l'inverso del tempo di ciclo.

Per le altre vale che

$$T_N = T_A + n/R$$

dove

- $T_N$  è il tempo medio per leggere o scrivere  $n$  bit
- $T_A$  è il tempo medio di accesso
- $n$  è il numero di bit
- $R$  è il tasso di trasferimento (in bit per secondo, o *bps*).

# Alterabilità

**Vi sono memorie il cui contenuto può essere scritto una volta sola (*Read Only Memory o ROM*).**

**In alcuni casi il contenuto di una memoria ROM può essere modificato off-line (*Programmable ROM o PROM*).**

# **Durevolezza del contenuto:**

## **Destructive Readout**

**Vi sono alcune tecnologie particolari (ad esempio le RAM dinamiche), nelle quali l'operazione di lettura causa la cancellazione del dato memorizzato (*Destructive Readout o DRO*).**

**In tal caso dopo ogni lettura è necessario eseguire un'operazione di riscrittura del dato.**

**Quindi in questo caso il tempo di ciclo è maggiore del tempo di accesso.**

# **Durevolezza del contenuto: refreshing**

- **In alcune tecnologie (ad esempio quelle delle memorie dinamiche) dopo un certo tempo i bit a 1 si trasformano in 0**
- **Questo effetto si ha ad esempio quando il bit è memorizzato sotto forma di carica all'interno di un condensatore, a causa delle correnti di scarica**
- **È quindi necessario che periodicamente si provveda a leggere ogni bit e a riscrivere i bit con valore 1 (*refreshing*).**

# **Durevolezza del contenuto: volatilità**

**Alcune memorie, come quelle RAM, perdono il loro contenuto quando non sono alimentate (memorie *volatili*).**

# Affidabilità

Le memorie possono essere colpite da due tipi di guasto:

- *Guasti transitori*: uno o più bit cambiano valore ad un certo istante, ma la memoria continua a funzionare correttamente
- *Guasti permanenti*: qualcosa nella memoria smette definitivamente di funzionare.

L'affidabilità è di solito misurata attraverso i seguenti parametri:

- *Mean Time To Failure (MTTF)*: tempo medio prima di avere un guasto
- *Failure rate (o Mean Time Between Failures, MTBF)*: frequenza media di occorrenza dei guasti.

# **Altre caratteristiche**

- **Tipo della memoria (elettronica, magnetica, meccanica, ottica)**
- **Consumo**
  - **può comportare la necessità di sistemi di raffreddamento**
  - **può essere critico per i sistemi portabili**
- **Portabilità**
- **Robustezza (ad esempio rispetto alle sollecitazioni meccaniche)**
- **Dimensione: dipende dalla densità di immagazzinamento.**