Le Memorie

Architettura del processore

Secondo il modello di Van Neumann la memoria dell'elaboratore è quel componente che serve ad immagazzinare informazioni (dati/istruzioni). Si tratta quindi di uno spazio dove è possibile memorizzare e leggere dei dati.

La memoria ideale

La memoria ideale ha:

- Capacità grande a piacere.
- Tempi di accesso alla memoria (sia in lettura che in scrittura) piccoli a piacere;
- Costi ridotti ed indipendenti dalla tecnologia impiegata.

La memoria reale

Capacità

La capacità di una memoria reale consiste nella sua dimensione misurata in Byte. Bisogna tener conto di:

- costo per bit: varia a seconda della tecnologia utilizzata;
- dimensione della word [bit]: quanti bit riesce a tirare fuori in un unico accesso;
- unità indirizzabile;
- unità di trasferimento.

Tempi

Si definisce *tempo di accesso* di una memoria reale il tempo impiegato per eseguire un ciclo completo di lettura/scrittura di un dato di assegnata grandezza.
Si definisce *tempo di durata del ciclo di memoria* la somma del tempo di accesso e di qualsiasi altro tempo necessario prima di un secondo accesso.
Si definisce *frequenza di trasferimento,* la frequenza alla quale i dati possono essere trasferiti da/verso la memoria.

Metodi di accesso

Per una memoria reale si può avere uno tra seguenti metodi di accesso:
Sequenziale (es: nastro magnetico):
L'accesso avviene in un dato ordine;

- C'è un'unica unità di lettura/scrittura;

Il tempo di accesso dipende dalla posizione del dato e dalla posizione a cui si è acceduto precedentemente.
Diretto (es: disco):
Ciascun blocco ha un indirizzo univoco basato sulla locazione fisica;
L'accesso è costituito da un salto al blocco seguito da una ricerca sequenziale;

- C'è un'unica unità di lettura/scrittura;
- Il tempo di accesso dipende dalla posizione del dato e dalla posizione a cui si è acceduto precedentemente.

Casuale (es: memoria centrale):
Ciascuna locazione ha un indirizzo univoco ed è raggiungibile direttamente;
Il tempo di accesso non dipende dalla posizione del dato e dalla posizione a cui si è acceduto precedentemente.
Associativo (es: memoria cache):
Ciascuna locazione ha un indirizzo univoco ed è raggiungibile direttamente;
La memoria è interrogata per contenuto, oltre che tramite un indirizzo;
Il tempo di accesso non dipende dalla posizione del dato e dalla posizione a cui si è acceduto precedentemente.

Caratteristiche fisiche
Per quanto concerne le caratteristiche fisiche, una memoria reale può essere:

- volatile: le informazioni decadono con il passare del tempo o quando manca l'alimentazione;
- non volatile: le informazioni permangono finché non sono volontariamente modificate e non richiedono energia elettrica;
- statica: memoria volatile capace di mantenere il proprio stato durante tutto il periodo in cui è alimentata;
- dinamica: memoria volatile capaci di mantenere il proprio stato solo per un periodo breve. Essa anche se alimentata necessita di periodici refresh. Un esempio è la RAM.
- riscrivibile: memoria che permette di cancellare e riscrivere più volte le informazioni;

non riscrivibile:

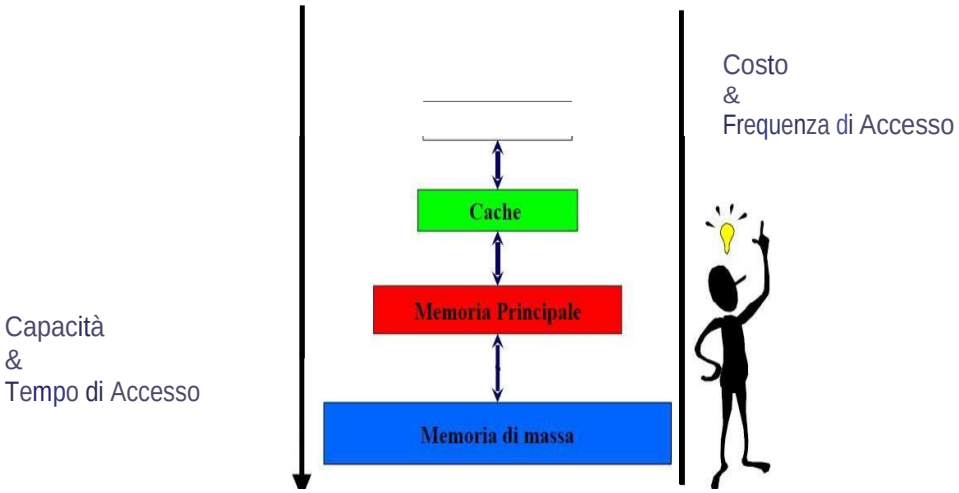
- ROM (Read Only Memory): i dati sono inseriti in memoria al momento della fabbricazione e non sono più modificabili;
- PROM (Programmable ROM): programmabili dall'utente, processo irreversibile;
- EPROM (Erasable PROM): cancellabili con luce ultravioletta e riprogrammabili;
- EEPROM (Electrically EPROM) :cancellabile e riscrivibili elettricamente;
- Flash.

Gerarchia di memorie

Le memorie reali sono caratterizzate da soluzioni legate a specifiche tecnologie ed associate a costi. Infatti, si verificano le seguenti relazioni:

- minore è il tempo di accesso, maggiore è il costo per bit;
- maggiore è la capacità, più basso è il costo per bit;
- maggiore è la capacità, più basso è il tempo di accesso;

Dato che è impossibile realizzare una memoria ideale la soluzione all'interno di un calcolatore è quella di non affidarsi ad un'unica memoria o tecnologia ma al contrario di utilizzare una gerarchia di memorie.
Osserviamo la figura.
Integrati nel processore ci sono i registri interni.
Nello stesso processore ma anche fuori abbiamo la memoria cache, la quale essendo una memoria statica, cioè che non ha bisogno di refresh, ed essendo molto vicina al processore è velocissima. Essa ha un costo molto alto (servono 6 transistor per una singola cella di memoria) per cui ha una capacità molto limitata. Ci possono essere vari livelli di cache a seconda della necessità.
Poi c'è la memoria principale (RAM), la quale è una memoria dinamica di dimensioni maggiori rispetto alla cache in quanto ha un costo più basso (serve 1condensatore per una singola cella di memoria), ma con un tempo di accesso minore.
Infine, c'è la memoria di massa che è permanente e riscrivibile e lenta.



Principio di località

Solitamente i programmi tendono a riutilizzare entro breve tempo, istruzioni/dati utilizzati di recente oppure vicini (in termini di indirizzi) ad una istruzione/dato eseguita di recente. Questo si verifica perché essi contengono molti cicli iterativi, subroutine ne ed effettuano operazioni su

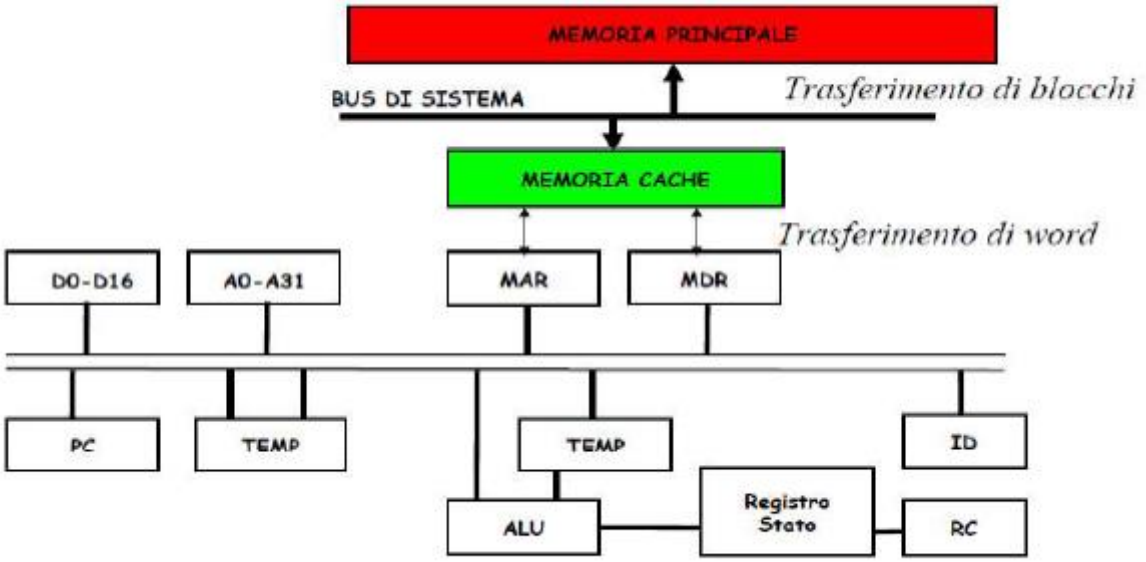
tabelle ed array.
Una regola empirica afferma che un programma trascorre, in media,l'80% del suo tempo di esecuzione in una porzione di codice, che rappresenta circa il 20% dell'intero programma. La medesima regola vale anche per i dati.

Quando si progetta un calcolatore bisogna garantire che le memorie più veloci siano quelle a cui si accede più spesso, mentre le memorie più lente contengano istruzioni/dati che saranno letti/scritti con bassa probabilità.
Per fare ciò ci si affida ad uno dei due seguenti principi:

- Principio di località temporale: esiste un'alta probabilità, che un'istruzione eseguita di recente sia nuovamente eseguita nell'immediato futuro.
  Tale principio chiede quindi di spostare una istruzione/dato nella memoria più veloce (cache), quando è richiesto per la prima volta, in modo che rimanga a disposizione nel caso di una nuova richiesta.
- Principio di località spaziale: esiste un'alta probabilità, che istruzioni vicine (in termini di indirizzi) ad un'istruzione eseguita di recente siano eseguite nell'immediato futuro.
  Tale principio chiede quindi di spostare nella memoria più veloce (cache) un insieme (blocco) di istruzioni/dati contigui a quello richiesto.

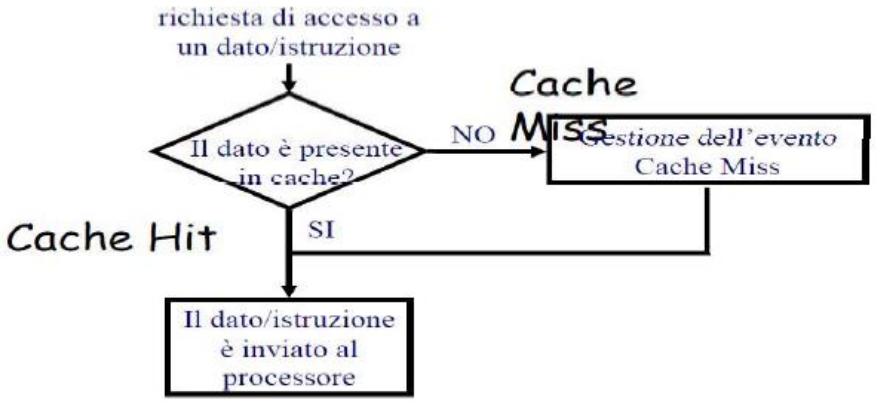
Memoria cache

La memoria cache è vicinissima al processore ed è collegata alla memoria principale. Lo scambio tra la memoria principale e la cache avviene a blocchi per cui il bus potrebbe essere abbastanza largo. Dalla memoria cache verso il processore (MAR e MDR) si ha un trasferimento di word.



Funzionamento

Il processore richiede un'istruzione o un dato alla cache. Se l'informazione è presente in cache viene inviata al processore (Hit, successo), altrimenti (Miss, insuccesso) deve essere gestito l'evento Cache Miss.



Quando si verifica un Cache Miss bisogna ricercare l'informazione richiesta dal processore nella RAM e verificare se nella cache c'è spazio libero. Se quest'ultimo c'è, viene copiato il blocco
Interessato dalla RAM alla cache, altrimenti è necessario capire quale blocco della cache riportare in RAM e sostituirlo con quello opportuno.

Parametri Caratteristici

I parametri che caratterizzano una memoria cache sono i seguenti:

- *Cache size:* dimensione della memoria cache [Byte];
- *Block Size:* dimensione di un blocco.

Un blocco è un insieme di locazioni di memoria contigue di una dimensione fissata;
La memoria cache può memorizzare solo un piccolo sottoinsieme di blocchi presenti in memoria principale;
Il principio di località spaziale suggerisce di spostare in cache un blocco invece che una singola istruzione/dato.

- *Hit Time:* numero di cicli di dock necessari per caricare un elemento che si trova in cache (in genere coincide con 1ciclo di dock);
- *Miss Penalty:* numero di cicli di dock necessari per caricare l'elemento dalla memoria principale nella memoria cache (solitamente è 10 volte);
- *Access Time:* tempo di accesso alla memoria principale;
- *Transfer Time:* tempo di trasferimento del blocco, in cui è presente l'elemento referenziato, dalla memoria principale alla memoria cache.
- *Miss Rate:* percentuale di miss che si verificano quando si richiede un'istruzione/dato alla memoria cache.

Funzione di Mapping

La *funzione di mapping* specifica la corrispondenza tra **i** blocchi della memoria principale e quelli della memoria cache.
In base alla funzione di mapping utilizzata, si possono distinguere 3 tipologie di memorie cache:

- Direct Mapped;
- Fully Associative;
- Set Associative.

Direct Mapped

La funzione di mapping *Direct Mapped* prevede che il dato in memoria centrale (presente in un blocco di un certo gruppo) può essere posizionato solo in un particolare blocco della memoria cache.
Uno dei vantaggi sta nel fatto che una volta calcolato l'indirizzo, è noto il blocco di cache che potrebbe contenere il dato, il quale può essere letto ed il processore può continuare a lavorare con esso prima che finisca di controllare che l'etichetta combaci effettivamente con l'indirizzo richiesto.

Lo svantaggio si ha quando ad esempio in memoria cache vi è il blocco zero di un gruppo e per il resto è vuoto. Nonostante la cache sia quasi vuota se serve il blocco zero di un altro gruppo è necessario compiere un'operazione di sostituzione.