La memoria

- Due tipologie:

- Memoria centrale
    - contiene i programmi in esecuzione e i relativi dati, la velocità di accesso incide fortemente sulle prestazioni del sistema.

- Memoria di massa
    - destinata a contenere grandi quantità di dati che non vengono utilizzati frequentemente, ma che devono essere mantenuti in modo persistente.

Criteri di valutazione

- Velocità di accesso
- Capacità
- Volatilità
- Costo per bit

Tipi di memoria

- La realizzazione delle unità di memoria si basano su diverse tecnologie:
  - memorie elettroniche
  - memorie magnetiche
  - memorie ottiche

Memorie elettroniche

- sono caratterizzate
  - da un'alta velocità (d'accesso ai dati)
  - da una discreta capacità
  - dalla necessità di essere continuamente alimentate (memorie volatili)
  - da un alto costo per bit

Memorie magnetiche

- Sono caratterizzate
  - da un basso costo per bit
  - da una bassa velocità di accesso
  - da grandi capacità
  - dalla capacità di mantenere le informazioni in assenza di alimentazione (memorie non volatili)

Memorie ottiche

- hanno proprietà analoghe a quelle delle memorie magnetiche
- sono in genere caratterizzate da supporti non riscrivibili
- adatte a memorizzare grandi quantità di dati, ma non alla memorizzazione di dati da elaborare.
- esistono dispositivi magneto-ottici, che consentono di modificare le informazioni memorizzate.

Memorie centrali e di massa

- Memorie centrali => basate su tecnologie elettroniche. Sono veloci, volatili, non molto grandi e costose
- Memorie di massa => basate su tecnologie magnetiche ed ottiche. Sono lente, molto capaci, poco costose e non volatili.
- Allo stato attuale
  - le memorie centrali sono 5 ordini di grandezza più veloci delle memorie di massa
  - le memorie centrali sono due ordini di grandezza più costose delle memorie di massa

La memoria centrale

- Può essere vista come una successione di elementi binari (bit) che sono in grado di assumere solo due stati (convenzionalmente, 0 o 1).

- I bit sono raggruppati in unità minime di 8, 16, 32 o 64 bit, dette celle.

- Ogni sequenza di bit avente la lunghezza di una cella è detta parola.

- Ogni cella possiede un indirizzo che rappresenta la sua posizione rispetto alla prima cella, la quale convenzionalmente ha indirizzo 0.

- la dimensione massima della memoria indirizzabile dipende dipende dal numero di bit a disposizione per codificare gli indirizzi di cella. Ossia, k bit per cella => 2^k indirizzi possibili.

Metodi di accesso

- Dato l'indirizzo della cella da leggere o scrivere, si possono individuare le seguenti tipologie di accesso:

  - Accesso sequenziale: le celle sono poste in successione. La lettura di un dato comporta la lettura di tutti i dati precedenti. (Esempio: nastri magnetici). Il tempo di accesso è molto variabile, dipende dalla posizione del dato nel supporto.

  - Accesso casuale: l'accesso a una cella non richiede la lettura delle precedenti. Il tempo d'accesso è indipendente dalla posizione del dato e può essere considerato costante. La memoria centrale è caratterizzata da un accesso casuale, è anche detta memoria RAM (Random Access Memory).

Metodi di accesso

- Accesso misto: L'indirizzo di una cella non consente di identificare con precisione la posizione fisica del dato sul supporto. Vengono effettuati una serie di accessi per giungere in prossimità del dato e successivamente una ricerca sequenziale per identificare con precisone la posizione del dato. Esempio: unità a dischi. Tempo di accesso variabile.

- Accesso associativo: Il contenuto di una cella non è selezionato in base all'indirizzo, ma in base a parte del contenuto della cella stessa. E' una sorta di accesso di tipo casuale, infatti il tempo di accesso è costante. Esempio: memoria cache. Sono memorie molto veloci e particolarmente costose.

Organizzazione della memoria

- La memoria è divisa su almeno due livelli.
  - Memoria di livello superiore: piccola, veloce e costosa
  - Memoria di livello inferiore: grande, lenta e a buon mercato
    - La memoria di livello superiore contiene un sottoinsieme dei dati e dei programmi residenti nella memoria di livello inferiore.
- Quando si cerca una parola, si cerca prima nella memoria di livello superiore e poi nella memoria di livello inferiore. Nel caso si trovi nel primo livello, l'accesso al dato sarà rapido, in caso contrario lento.
- Un organizzazione di questo genere è efficiente solo se esiste un buon criterio per scegliere quali dati devono essere trasferiti dalla memoria lenta a quella veloce.

Organizzazione della memoria

- È necessario prevedere con una certa precisione il comportamento dei programmi.

- Statisticamente, un programma indirizza il 90% delle richieste di lettura/scrittura a un'area di memoria contigua inferiore al 10% dell'area complessiva.

  - Località spaziale: quando un programma fa riferimento a un dato, è molto probabile che faccia riferimento a un dato fisicamente vicino nell'immediato futuro.

  - Località temporale: quando un programma fa riferimento a un dato, è molto probabile che si riutilizzi nell'immediato futuro.

Principio di località

- Principio di Località: le informazioni usate di recente e quelle in posizioni contigue saranno utilizzate con molta probabilità nell'immediato futuro.
- Se si favorisce l'accesso a tali informazioni (precaricandole nella memoria di livello superiore), si migliorano le prestazioni del sistema.
- Se la memoria di livello superiore è già piena (di solito è di piccole dimensioni), le informazioni che soddisfano il principio di località vanno a sostituire un blocco di memoria di livello superiore, i cui dati non sono stati recentemente utilizzati (Least Recently Used Policy, LRU).

Organizzazione reale

- Generalmente ci sono più di due livelli.

Registri

⇩

Memoria Cache

⇩

Memoria RAM

⇩

Dischi magnetici

⇩

Nastri

Dispositivi di memoria di massa

- Nastri magnetici
  - sono dispositivi ad accesso sequenziale
  - le operazioni di lettura/scrittura sono effettuate mediante una testina che può rilevare e modificare lo stato di magnetizzazione del nastro
  - informazioni scritte e lette a blocchi chiamati record fisici, separati tra loro da spazi chiamati gap.
  - Dispositivi economici, lenti, ma di grandi capacità. Utili per i backup.

Dispositivi di memoria di massa

- Dischi magnetici
  - sono dispositivi ad accesso misto
  - l'informazione memorizzata in anelli concentrici detti tracce. In ogni traccia è memorizzata la stessa quantità di informazione (=> densità variabile di memorizzazione). Ogni traccia è suddivisa in settori. Settori e tracce sono separati da gap.
  - per poter utilizzare un disco, si deve dapprima organizzarlo in settori e tracce mediante l'operazione di formattazione.
  - per leggere e scrivere dei dati è necessario fornire alla testina numero di traccia e numero di settore.
  - tipiche unità a dischi: floppy disk, hard disk.

Dispositivi di memoria di massa

- Dischi ottici
  - sono generalmente dispositivi non riscrivibili (CD-ROM)
  - permettono di memorizzare grandi quantità di dati.
  - i dati sono memorizzati rendendo opache o lucide le zone del disco (0 o 1). La lettura avviene mediante un raggio laser che esplora la superficie e identifica il valore dei bit in base alla riflessione o meno del raggio. L'informazione è organizzata in un unico percorso a spirale. => accesso sequenziale, che comunque risulta essere piuttosto veloce.
  - esistono anche dischi ottici che permettono di essere riscritti più volte: CD-RW, DVD-RW (Rewritable)
  - DVD: dischi ottici di ultima generazione, molto piu' capienti dei CD.
    .

Interfacce di I/O

- il calcolatore comunica con l'ambiente esterno (le periferiche) mediante delle interfacce di ingresso/uscita, che hanno il compito di tradure i segnali che giungono dal calcolatore in informazioni comprensibili alle periferiche e vice versa.
- La trasmissione dei dati può essere
  - seriale (l'informazione è trasmessa un bit per volta)
  - parallela (più bit trasmessi in parallelo)
- Alcuni standard
  - SATA, USB, FireWire (tx seriale)
  - Centronics (tx parallela), PATA

Interfacce di I/O

- Ogni interfaccia di I/O è dotata almeno dei seguenti registri
  - Registro dati, utilizzato per scambiare i dati tra periferica e calcolatore. Connesso con il bus dati.
  - Registro di stato (o di controllo), nel quale transitano informazioni di controllo necessarie alla sincronizzazione tra CPU e periferica. Connesso con il bus di controllo.

Sincronizzazione

- Periferiche e CPU hanno generalemnte diverse velocità e necessitano di sincronizzazione.
- Ci sono tre diversi metodi di sincronizzazione:
  - a controllo di programma
  - a interruzione
  - con accesso diretto alla memoria

Sync. a controllo di programma

- La sincronizzazione è completamente gestita dalla CPU.
- La CPU esegue un ciclo (detto ciclo di polling) che ispeziona/scrive periodicamente il registro dati.
- Esempio: stampa di una linea di caratteri mediante una stampante, ogni singolo carattere viene trasferito alla stampante mediante il registro dati, solo quando un carattere è stato stampato viene trasferito nel registro dati il seguente.
- Vantaggio: è una gestione della sincronizzazione semplice.
- Svantaggio: rischio di sovraccarico della CPU

Sync. a interruzione

- Elimina il problema di sovraccarico della CPU tipico della sincronizzazione a controllo di programma.

- Ogni interfaccia è dotata della possibilità di notificare il suo status alla CPU mediante un segnale chiamato interruzione (o interrupt).

- Quando la CPU riceve un interrupt, interrompe la sua attività ed esegue un programma di risposta all'interruzione per gestire la comunicazione con l'interfaccia (e quindi con la periferica).

- La CPU è occupata solo per il trasferimento dei dati. (si evitano i tempi di attesa del ciclo di polling)

Sync. con accesso diretto alla memoria

- Se si hanno grossi e frequenti trasferimenti di dati, la gestione della sincronizzazione mediante interrupt rischia di essere inefficiente.

- Esistono delle componenti HW chiamate DMA (Direct Memory Access) che sostituiscono la CPU nella gestione del trasferimento dati.

- La CPU controlla i DMA, comunica loro solo l'indirizzo di memoria da cui leggere o sul quale scrivere e la quantità di dati da trasferire, dopodiché il DMA gestisce l'intero processo di trasferimento.

- Nelle architetture più sofisticate i DMA sono processori dedicati all'input/output.