FONDAMENTI DIINFORMATICA

Lezione n. 13

- MEMORIE VLSI, MEMORIE MAGNETICHE
- EVOLUZIONE, COSTI, CAPACITÀ, PRESTAZIONI
- PRINCIPIO DI LOCALITÀ
- CONCETTI DI BASE E TECNOLOGIA DELLE MEMORIE
- DEFINIZIONE DI HIT RATIO
- ANALISI DEI TEMPI DI ACCESSO GLOBALI

Nelle prossime lezioni esamineremo le caratteristiche e l'organizzazione delle memorie che hanno una influenza determinante sulle prestazioni dei sistemi di elaborazione.
Le prestazioni sono influenzate dalla tecnologia che sono in rapidissima evoluzione.

LA MEMORIA

I sistemi di memoria di un elaboratore possono essere suddivisi in:

- **Memoria interna al processore.**
- **Memoria principale.**
- **Memoria secondaria.**

LA MEMORIA INTERNA

- **Registri interni alla CPU**
  - **visibili o no al programmatore**
  - **memorizzano temporaneamente dati e istruzioni.**
  - **dimensioni: decine di bytes.**
  - **tempo di accesso: qualche ns.**

Nelle CPU più recenti cresce la quantità di risorse dedicate alla memoria:

- **memorie cache nella CPU:**
  - **1980: processori senza cache (I386)**
  - **1995: Alpha 21164 55% dei trans.**
  - **2000: Merced(Intel-HP) 85% dei trans.**

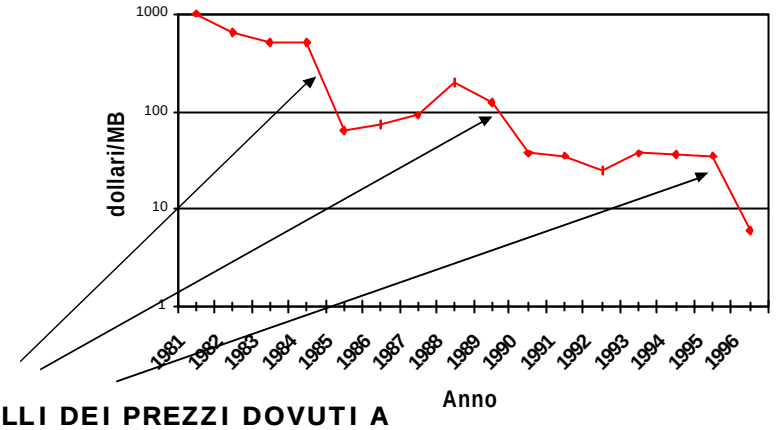
MEMORIA PRINCIPALE

- **Veloce e di grande capacità.**
- **Memorizza dati e istruzioni che servono per il funzionamento dell'unità centrale.**
- **La CPU vi accede direttamente.**
- **Dimensioni: decine di Mbytes.**

E' la memoria nella quale sono contenuti i programmi che la CPU esegue e i dati cui la stessa CPU può accedere direttamente.

Poche decine Mbytes su un personal computer, centinaia di MBytes su supercalcolatori.

Prezzo dollari/MB memoria DRAM



CROLLI DEI PREZZI DOVUTI A SOVRAPRODUZIONE

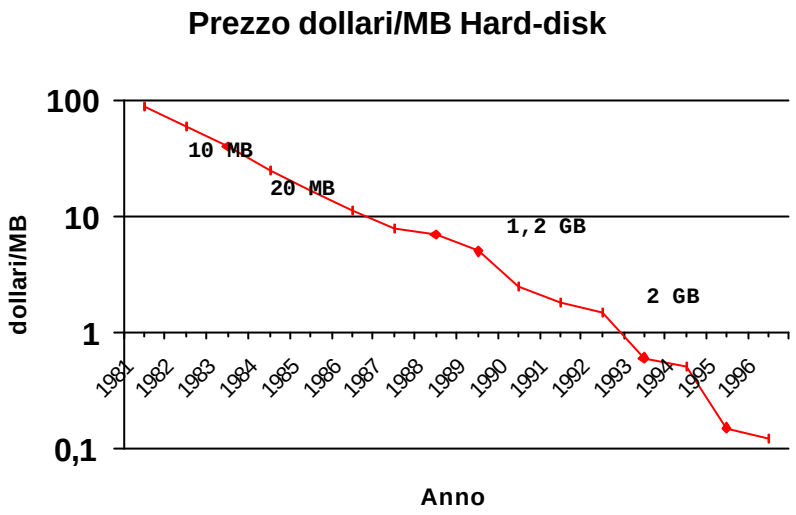
MEMORIA SECONDARIA

- **Di grandi dimensioni (Gbytes) e molto più lenta della memoria principale.**
- **Memorizza dati e istruzioni che non sono di immediato interesse della CPU.**

Può essere suddivisa in:

- **Memoria in linea (es. dischi magnetici). Interesse nell'ambito di millisecondi ... secondi.**
- **Memoria fuori linea (es. nastri magnetici). Interesse nell'ambito di minuti ... anni.**

I sistemi di memoria secondaria utilizzano ora le tecnologie sviluppate per applicazioni di largo consumo. Le tecnologie della riproduzione video o dei suoni ad alta fedeltà nell'ambito dei sistemi di elaborazione ha modificato il panorama tecnologico e ridotto i costi dei sistemi di memoria secondaria.



TECNOLOGIE E CARATTERISTICHE

I vari tipi di memoria sono realizzati con tecnologie con valori diversi di:

- **Costo per singolo bit immagazzinato.**
- **Tempo di accesso (ritardo fra l'istante in cui avviene la richiesta e l'istante in cui il dato è disponibile al richiedente)**
- **Modo di accesso (seriale o casuale).**

TECNOLOGIA DELLE MEMORIE

Memorie a semiconduttore con tecnologia VLSI (memoria principale).

Memorie magnetiche (memoria secondaria).

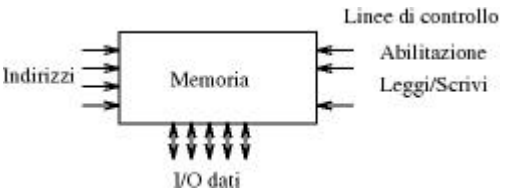
Memorie ottiche (memoria secondaria).

MEMORIE A SEMICONDUTTORE

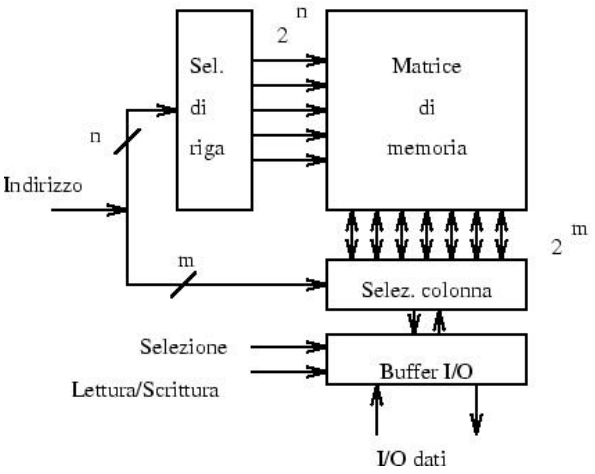
La tecnologia VLSI realizza su un circuito integrato memorie di capacità sempre crescenti.

In ogni circuito integrato sono contenute:

- **le celle di memoria,**
- **i circuiti di decodifica dell'indirizzo,**
- **le interfacce di uscita di potenza (buffer) e i circuiti di ingresso.**



MEMORIE A SEMICONDUTTORE



LE MEMORIE ROM

ROM - *Read Only Memory* o memorie a sola lettura.
La CPU, durante l'esecuzione di un programma, può effettuarne solo la lettura.
L'informazione permane anche se viene meno la tensione di alimentazione.

La scrittura può essere effettuata con modalità e tempi diversi:
PROM: *Programmable ROM* - La memoria è scrivibile, dal costruttore o dall'utilizzatore, una volta per tutte.
EPROM: *Erasable PROM* - La memoria è scrivibile all'utilizzatore e cancellabile con raggi ultravioletti.
EAROM: *Electrically Alterable ROM* - Le celle di memoria sono più volte riscrivibili elettricamente.

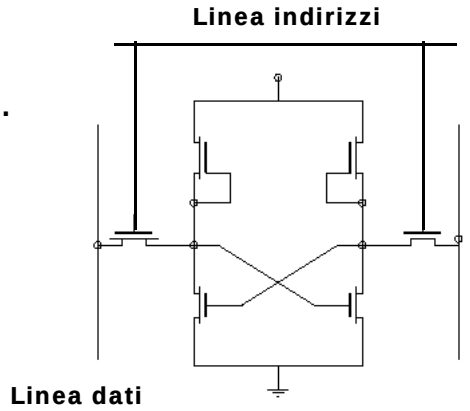
LE MEMORIE RAM

RAM - Random Access Memory.
- **RAM: (di solito) memorie a semiconduttore ad accesso casuale che sono sia leggibili sia scrivibili.**
- **L'informazione scompare se viene meno la tensione di alimentazione.**
- **RAM statiche o RAM dinamiche.**

L'acronimo RAM viene utilizzato correntemente per indicare le memorie a lettura e scrittura utilizzate come memorie principali di un sistema di elaborazione.

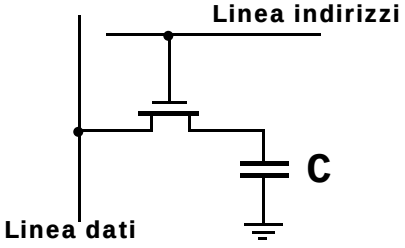
LE MEMORIE RAM STATICHE

- La cella elementare è costituita da 6 transistori MOS che formano un FLIP-FLOP.
- L'informazione permane stabile in presenza della tensione di alimentazione.
- Tempi di accesso rapidi.
- Costi elevati



LE MEMORIE RAM DINAMICHE

- La cella elementare è costituita da un condensatore che viene caricato (1) o scaricato (0).
- La tensione sul condensatore tende a diminuire (millisecondi) e quindi deve essere ripristinata o rinfrescata.



LE MEMORIE RAM DINAMICHE

- La semplicità della cella consente capacità molto elevate (milioni di bit).

Anno	Dimensioni	Tempo di ciclo
1980	64 Kbit	250 ns
1983	256 Kbit	220 ns
1986	1 Mbit	190 ns
1989	4 Mbit	165 ns
1992	16 Mbit	140 ns
1995	64 Mbit	120 ns

MEMORIE AD ACCESSO SERIALE

- Condividono il sistema (o testina) di lettura e scrittura tra diverse locazioni di memoria.
- La sequenza di locazioni che condivide la stessa testina si chiama *traccia*.
- L'accesso alla locazione di memoria avviene spostando la testina o la traccia.
- La traccia o parte di essa deve essere letta completamente per accedere al singolo dato.
- Le memorie ad accesso seriale hanno raggiunto con la tecnologia magnetica costi per bit estremamente competitivi.

MEMORIE AD ACCESSO SERIALE

Le memorie seriali hanno tempi di accesso elevati perché:

- Occorre tempo per posizionare la testina di lettura.
- La traccia si muove a velocità ridotta.
- Il trasferimento dati è seriale.
- La testina di lettura è condivisa fra più tracce.

TEMPO DI ACCESSO

- *seek time* (t_s): necessario alla testina per spostarsi da una traccia all'altra.
- *latency time* o *tempo di latenza* (t_L): necessario per posizionare la testina sul dato da leggere (o scrivere). Se r è la velocità di rotazione il tempo medio diventa:

$$t_L = (2r)^{-1}$$

TEMPO DI ACCESSO

Il tempo di lettura di un blocco di dati che dipende dalla velocità relativa fra la traccia e la testina di lettura.

tempo lettura di un dato = $(N)^{-1}(r)^{-1}$
dove N è la lunghezza della traccia.

ESEMPI

Disco	t_s [ms]	N[Kbytes]	r[giri/min]	t_L[ms]
NEC D2257 (1985)	20	20	3510	8,5
Quantum (1995)	7,9	74	7200	4,2

TEMPO DI ACCESSO

Tempo di accesso t_B ad un blocco di lunghezza n:
$t_B = t_s + t_L$ + (tempo di lettura blocco) =
$t_s + (2r)^{-1} + n(N)^{-1} (r)^{-1}$ =

Nel caso del NEC D2257
$(20 + 8,5 + n*8,5*10^{-3})$ms =
[28,5 + 8,5*(Kbytes)]ms

Nel caso del Quantum
$(7,9 + 4,2 + n*1,2*10^{-3})$ms =
[12,1 + 1,2*(Kbytes)]ms

IL SISTEMA DI MEMORIA

Le memorie di un calcolatore formano un sistema unico che deve essere progettato e gestito in modo da ottenere:

- **Capacità di memorizzazione adeguata.**
- **Prestazioni accettabili.**
- **Costi ridotti.**

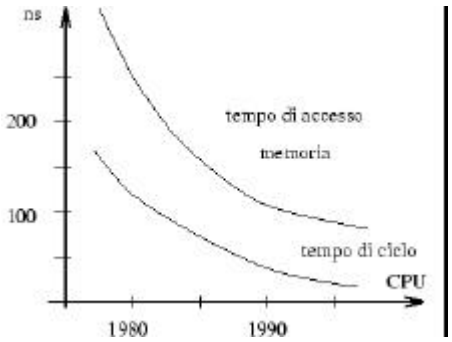
Gli obiettivi indicati sono ovviamente in contrasto fra loro.
Lo scopo del progetto architetturale è quello di raggiungere un ragionevole compromesso fra gli obiettivi indicati.

CPU-MEMORIA

Nell'architettura VonNeuman il canale di comunicazione tra la CPU e la memoria è il punto critico (collo di bottiglia) del sistema.

- **La tecnologia consente di realizzare CPU sempre più veloci.**
- **Il tempo di accesso delle memorie non cresce così rapidamente.**

CPU-MEMORIA



Sono disponibili nel 2000 CPU con frequenza di clock superiore a 500 MHz.

Le prestazioni delle CPU non devono essere troppo negativamente influenzate dal tempo di accesso alle memorie.

LA GERARCHIA DELLE MEMORIE

La soluzione ottimale per un sistema di memoria è:

- **Costo minimo.**
- **Capacità massima.**
- **Tempi di accesso minimi.**

Soluzione approssimata: GERARCHIA

Tecnologie diverse possono soddisfare al meglio ciascuno dei requisiti.
Una gerarchia cerca di ottimizzare globalmente i parametri.

ESEMPIO DI GERARCHIA

Il sistema di memoria di uno studente ha una struttura gerarchica:

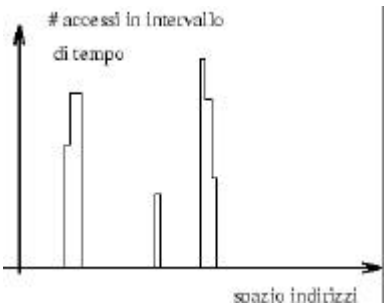
- **La propria memoria.**
- **La borsa.**
- **Lo scaffale di casa.**
- **La libreria o la biblioteca di Facoltà.**
- **Depositi casa editrice.**

La gestione del sistema di memoria globale di uno studente è molto complessa e richiede la conoscenza preventiva delle attività che si svolgeranno.

PRINCIPIO DI LOCALITA'

- **Un sistema di memoria gerarchico può essere reso efficiente se la modalità di accesso ai dati ha caratteristiche prevedibili.**
- **Il meccanismo di prevedibilità è il *Principio di località:***
  " Se al tempo t si accede all'indirizzo X è "molto probabile" che l'indirizzo X+DX sia richiesto fra t e t+ D t" .
- **Nel breve periodo gli indirizzi generati da un programma sono confinati in regioni limitate.**

LOCALITA'



La distribuzione degli accessi alla memoria in un dato intervallo può essere misurato direttamente su un sistema.

LA GERARCHIA

Coppia di strutture di memoria M1 e M2 con:

- **costo per bit: $c_1 > c_2$,**
- **dimensioni: $S_1 < S_2$,**
- **tempi di accesso: $t_{A1} < t_{A2}$**

- **M_1 e M_2 realizzati con tecnologie diverse.**
- **Gestione della gerarchia automatica e invisibile all'utente.**
- **Sono attualmente utilizzati sistemi con più livelli di gerarchia della memoria.**

CRITERI DI GESTIONE

- **I dati utilizzati più spesso vanno posti in memorie facilmente accessibili.**
- **I dati utilizzati più raramente sono posti in memorie con tempi di accesso elevato.**
- **Allocazione dinamica per utilizzare gli spazi disponibili con la massima efficienza.**
- **Spostamento automatico dei dati tra i livelli.**
- **Canali di comunicazione veloci fra i livelli.**

La politica di gestione tende a mimare una memoria che abbia:

- **i tempi di accesso della più veloce,**
- **le dimensioni della maggiore,**
- **i costi della più economica.**

HIT E MISS RATIO

Le prestazioni del sistema sono determinate dal:

tasso di successo o *Hit ratio* = H

definito come la probabilità che la richiesta sia soddisfatta al livello M_1.

Si definisce tasso di insuccesso o *Miss ratio* la probabilità che la richiesta non sia soddisfatta al livello M_1.

Miss ratio = 1 - H

TEMPO ACCESSO

Tempo di accesso medio globale:

$$t_A = H\, t_{A1} + (1- H) t_{A2}$$

dove

- $t_{A2} = t_{A1} + t_B = r\, t_{A1}$
- **t_B è il tempo di accesso a un blocco di M_2.**

Efficienza di accesso =

$$e = \frac{t_{A1}}{t_A} = \frac{t_{A1}}{H t_{A1} + (1-H) t_{A2}} = \frac{1}{H + (1-H) r}$$

dove: $r = t_{A2} / t_{A1}$

PRESTAZIONI