



UNIVERSITA' DEGLI STUDI DI PARMA

DIPARTIMENTO DI INGEGNERIA DELL'INFORMAZIONE

# **Fondamenti di Informatica B**

---

Lezione n. 13

**Alberto Broggi – Gianni Conte**

**A.A. 2005-2006**

# Fondamenti di Informatica B

---

## Lezione n.13

- MEMORIE VLSI, MEMORIE MAGNETICHE
- EVOLUZIONE, COSTI, CAPACITÀ, PRESTAZIONI
- CONCETTI DI BASE E TECNOLOGIA DELLE MEMORIE
- PRINCIPIO DI LOCALITÀ
- GERARCHIA DI MEMORIA
- MEMORIA VIRTUALE E MEMORIA CACHE

In questa lezione 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 è in rapidissima evoluzione.

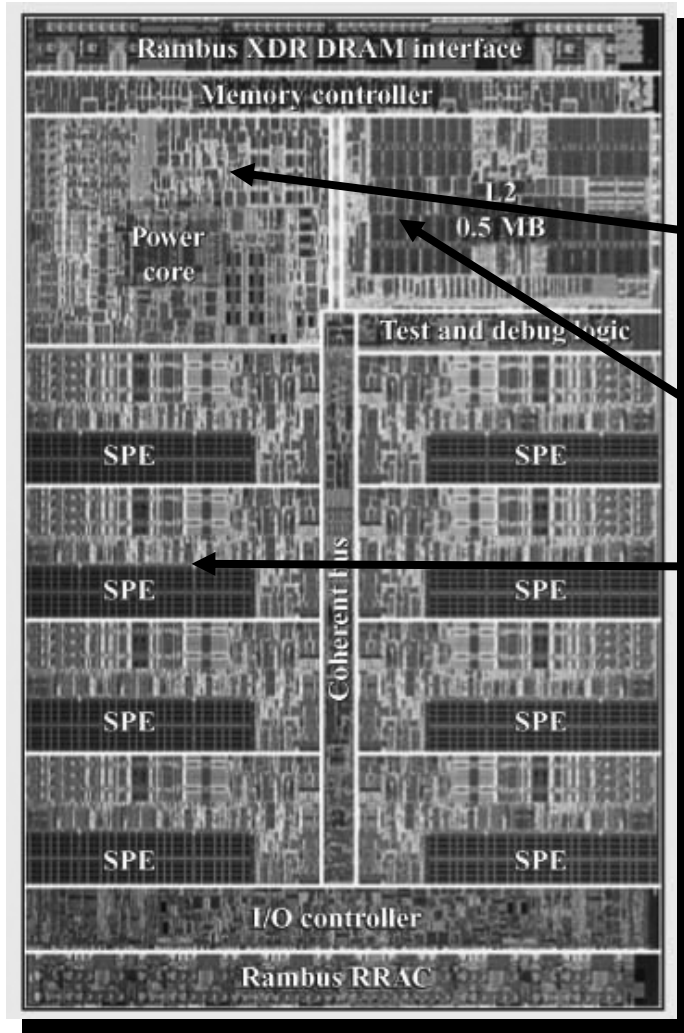
# Le Memorie

---

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

- Memoria interna al processore
- Memoria principale
- Memoria secondaria

# La Memoria Interna



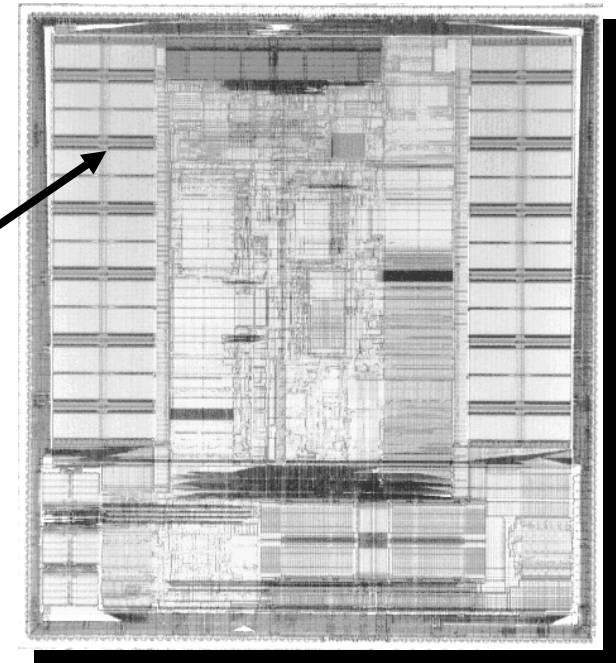
**Processore CELL (2005)**

Memoria sul chip  
del processore

Registri interni

Memorie cache

Memorie interne al  
processore



**Alpha 21164 (1995)**

18.1mm x 16.5mm  
contains 9.3 million transistors  
96 kB L2 cache

super-scalar architecture, capable of  
issuing 4 instructions per clock cycle

# La Memoria Interna

---

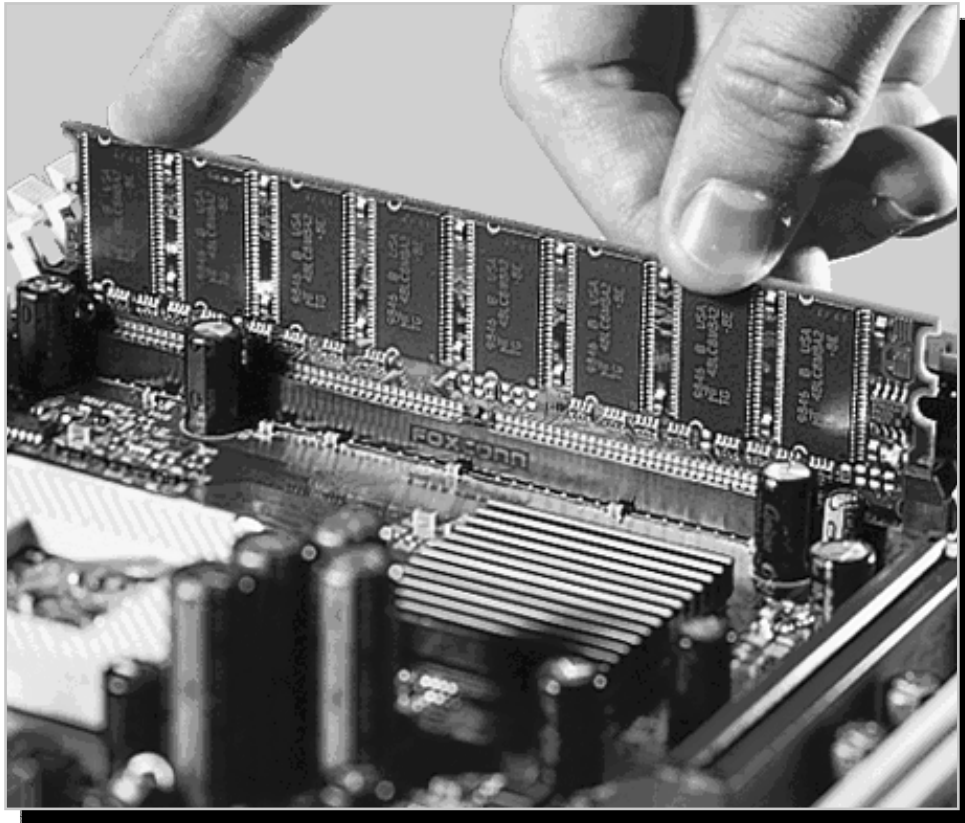
- Registri interni alla CPU
  - visibili o no al programmatore
  - memorizzano temporaneamente dati e istruzioni
  - dimensioni: decine di bytes
  - tempo di accesso: qualche ns

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 transistori
  - 2000: Merced (Intel-HP), 85% dei transistori

# La Memoria Principale

---



- Memoria sulla scheda madre
- Moduli SIMM (Single Inline Memory Module)

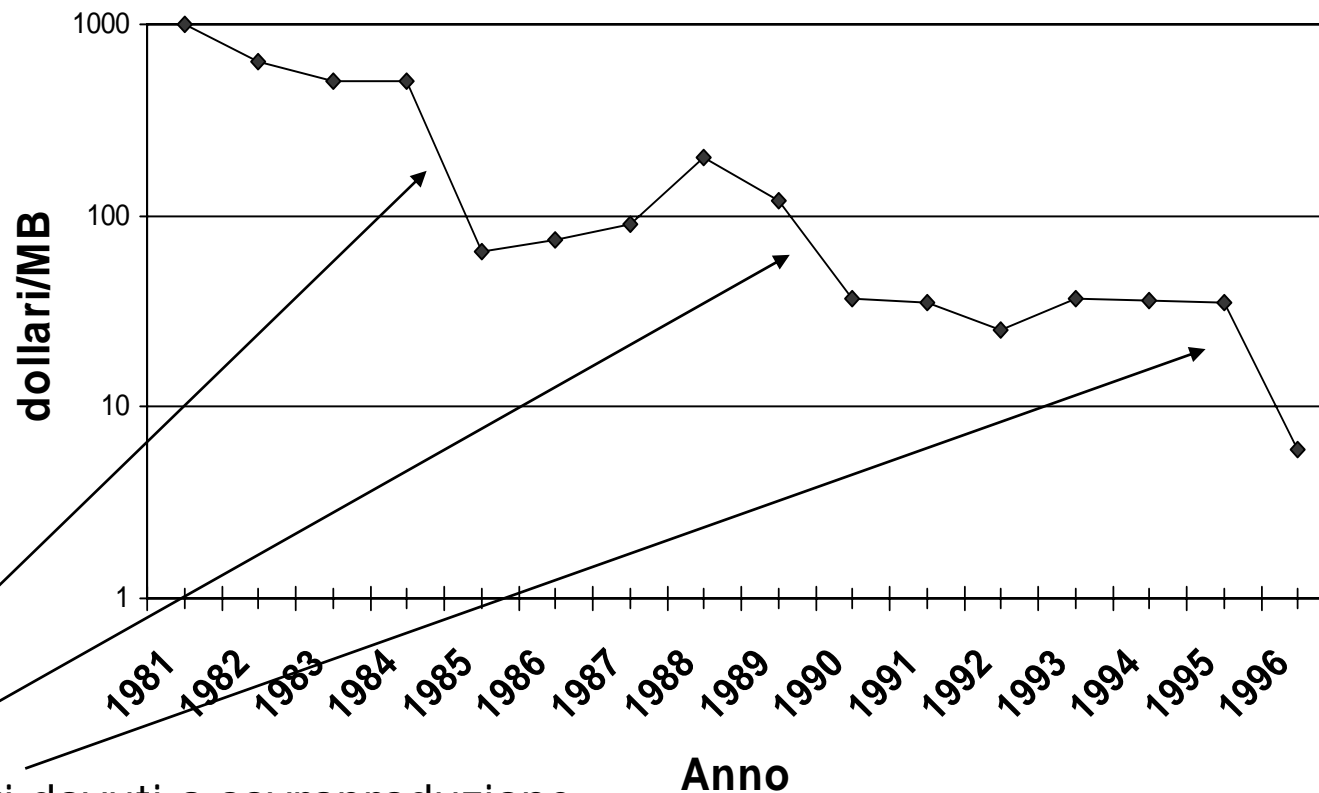
# La Memoria Principale

---

- Veloce e di grande capacità
- Memorizza dati e istruzioni che servono per il funzionamento dell'unità centrale
- La CPU vi accede direttamente
  
- Nella memoria sono contenuti i programmi che la CPU esegue e i dati cui la stessa CPU può accedere direttamente
  - Centinaia Mbytes su un personal computer
  - GBytes su server e supercalcolatori

# La Memoria Principale

## Prezzo dollari/MB memoria DRAM



Crollo dei prezzi dovuti a sovrapproduzione

**2006: 512 Mbyte circa 50 € e 0,1 dollaro/MB**



# La Memoria Secondaria

---



**Disco**

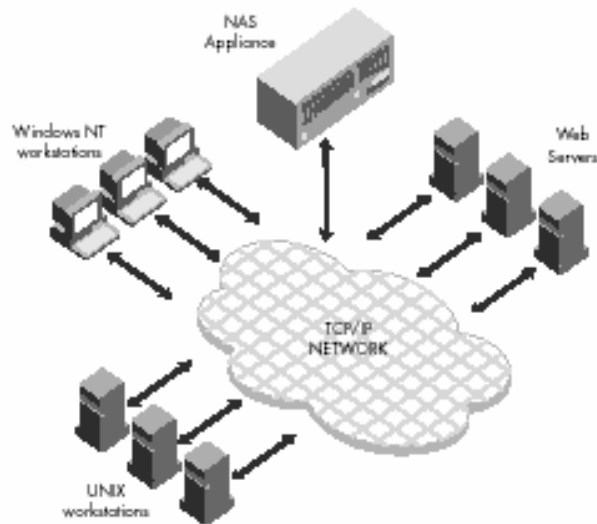


**NAS: Network Attached Storage**

**SAN: Storage Area Network**

**Sistemi fino a 100 Tbyte**

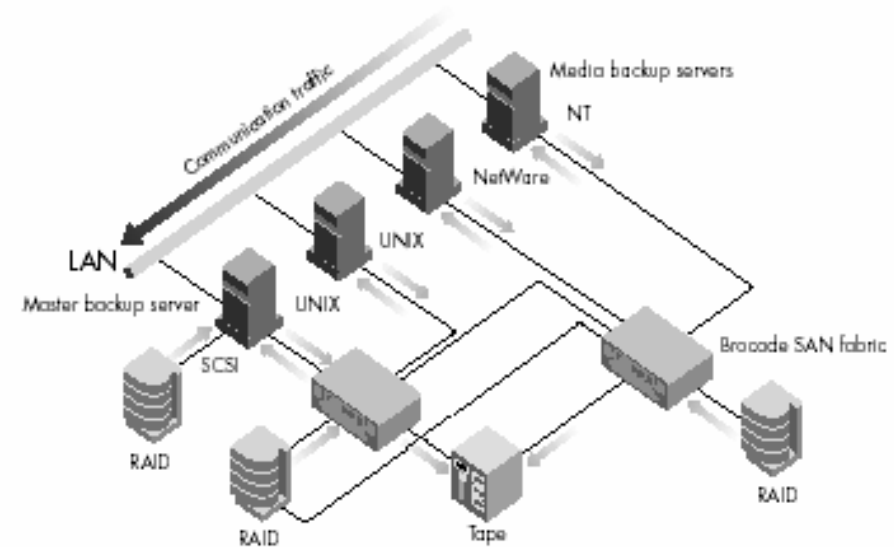
# La Memoria sulla Rete



## NAS

(Network Attached Storage)

Sistemi di memoria connessi al resto del sistema mediante la rete locale



## SAN

(Storage Area Network)

Sistemi di memoria connessi agli elaboratori mediante canali veloci a fibra ottica

# La 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

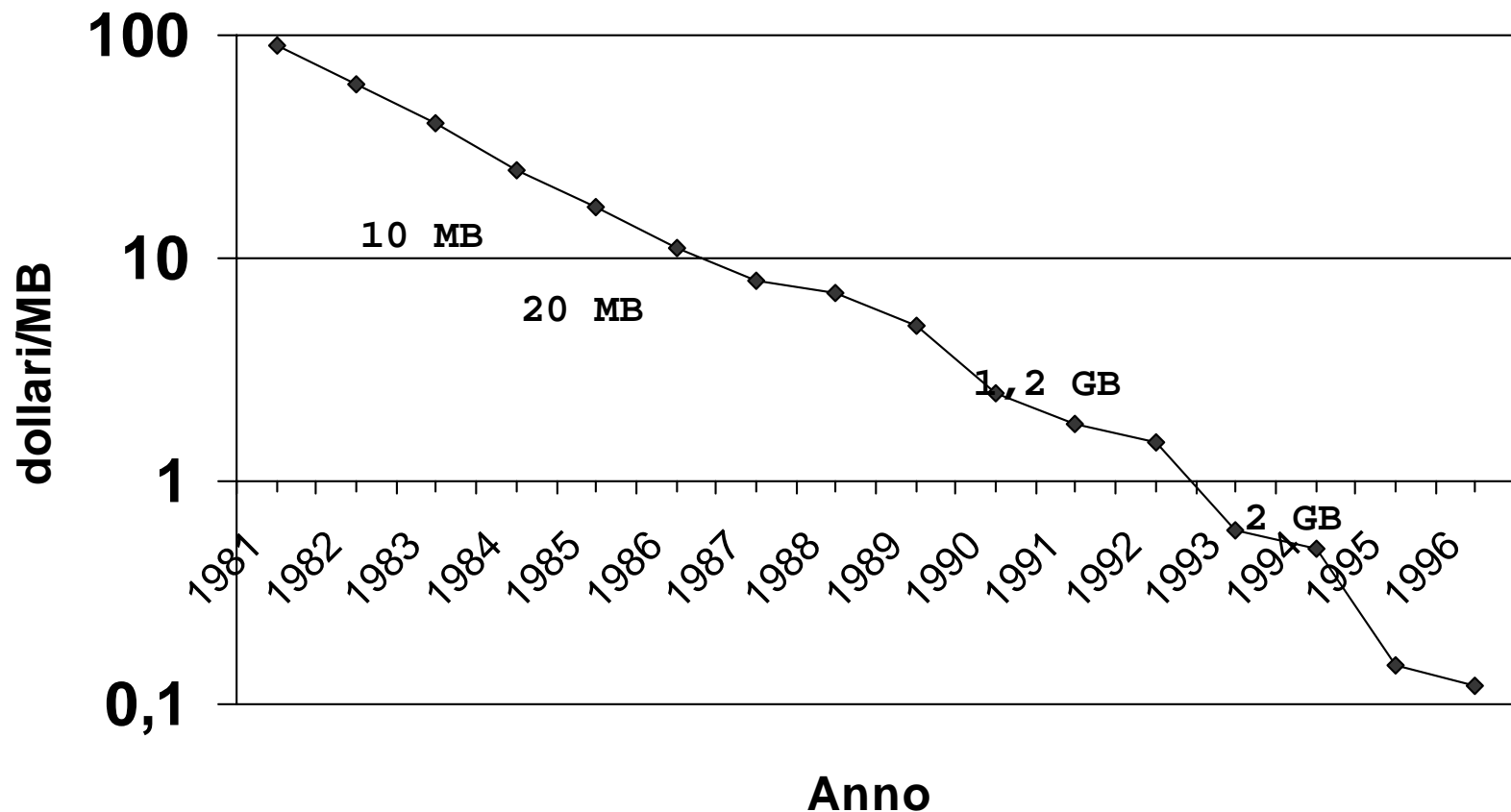
# La Memoria Secondaria

---

- 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

# La Memoria Secondaria

## Prezzo dollari/MB Hard-disk



2006: 200 Gbyte 100 € e 0,0005 dollaro/MB

# 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)

# **Tecnologie e Caratteristiche**

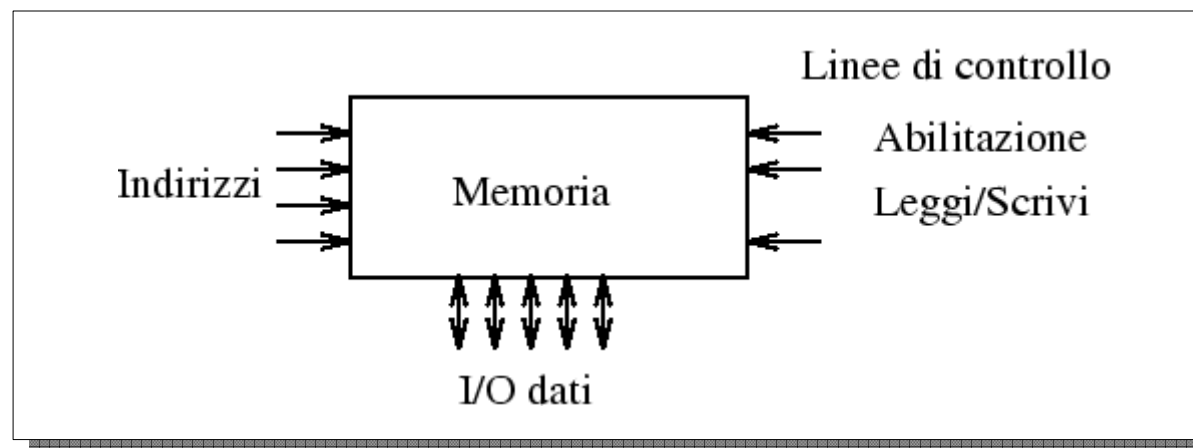
---

- 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 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 Sola Lettura

---

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

# Memorie a Sola Lettura

---

- 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* (EEPROM o E<sup>2</sup>PROM) - Le celle di memoria sono più volte riscrivibili elettricamente

# Memorie RAM

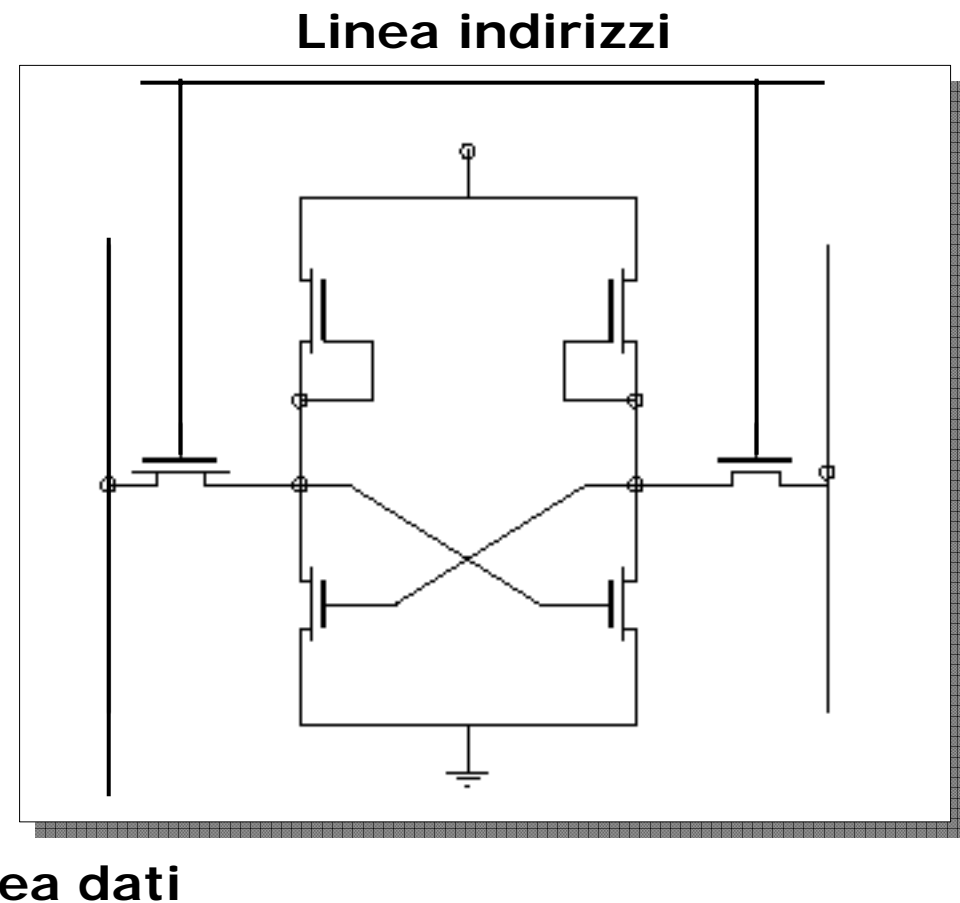
---

- RAM - Random Access Memory
- Memorie (di solito) 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

# Memorie RAM Statiche - SRAM

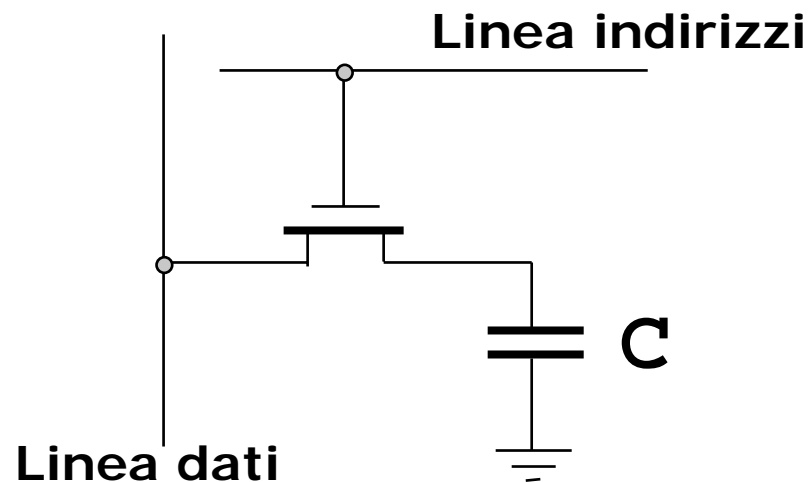
- La cella elementare è costituita da 6 transistori MOS che formano un FLIP-FLOP
- L'informazione permane stabile in presenza della tensione di alimentazione
- Tempi di accesso rapidi
- Costi elevati



# Memorie RAM Dinamiche - DRAM

---

- 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



# Memorie RAM Dinamiche

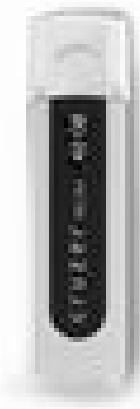
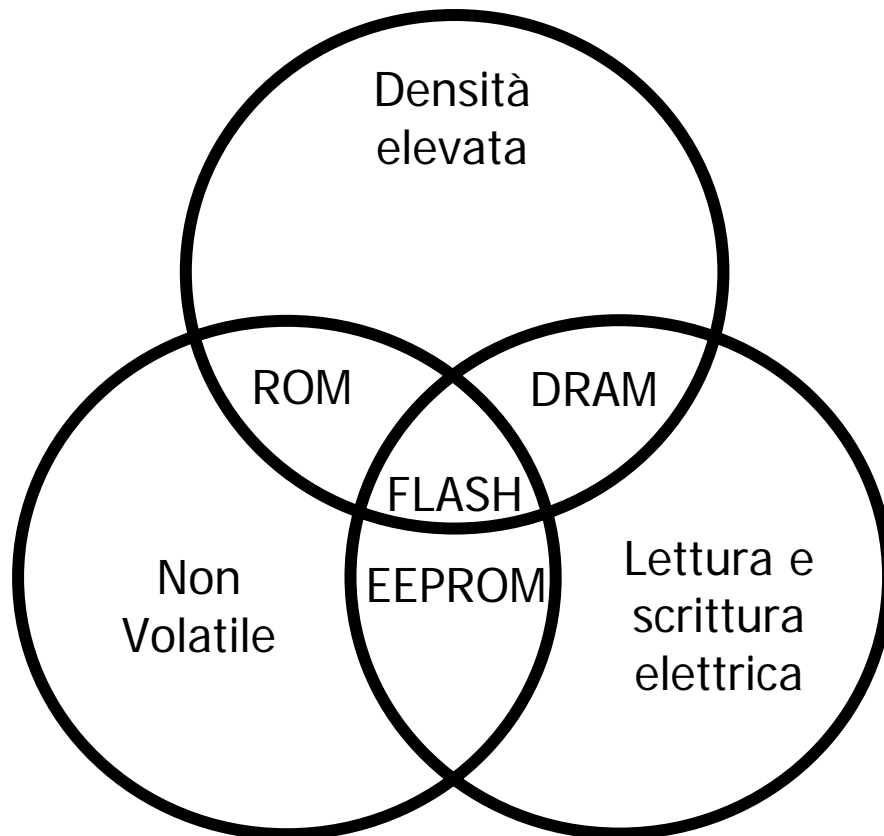
---

- 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         |
| 2000 | 1 Gbit     | 70 ns          |
| 2006 | 4 Gbit     | 50 ns          |

# Memorie Flash

Le Flash memory sono memorie a stato solido (non hanno parti in movimento) e sono memorie non volatili (mantengono l'informazione anche se non sono alimentate) che si stanno diffondendo in molteplici campi



# Memorie ad Accesso Seriale

---

- Nastri magnetici, dischi magnetici e ottici:
  - Condividono il sistema di lettura (o testina) 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 e ottica costi per bit estremamente bassi



# Dischi Magnetici

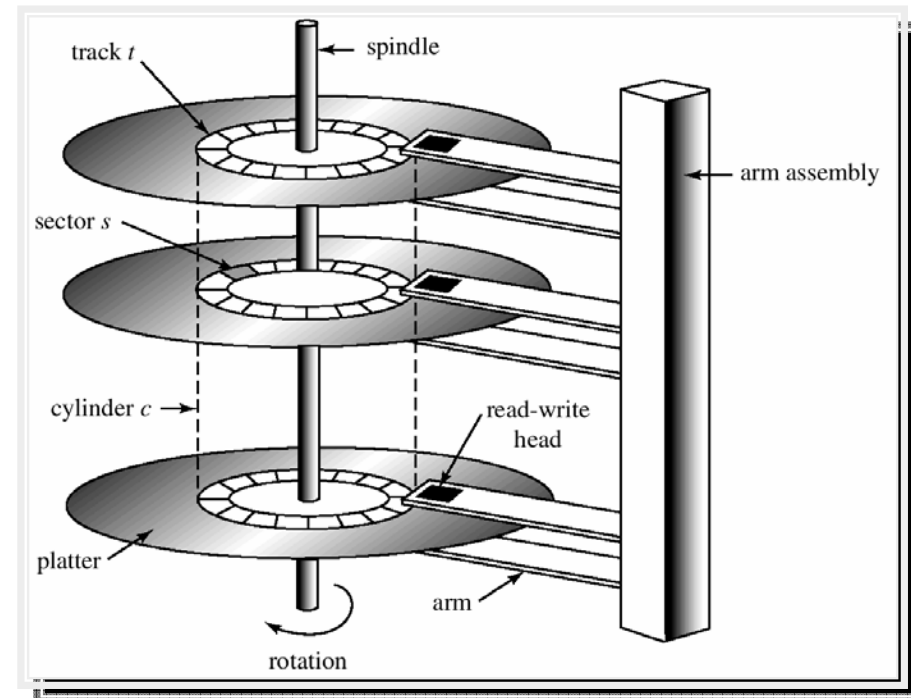
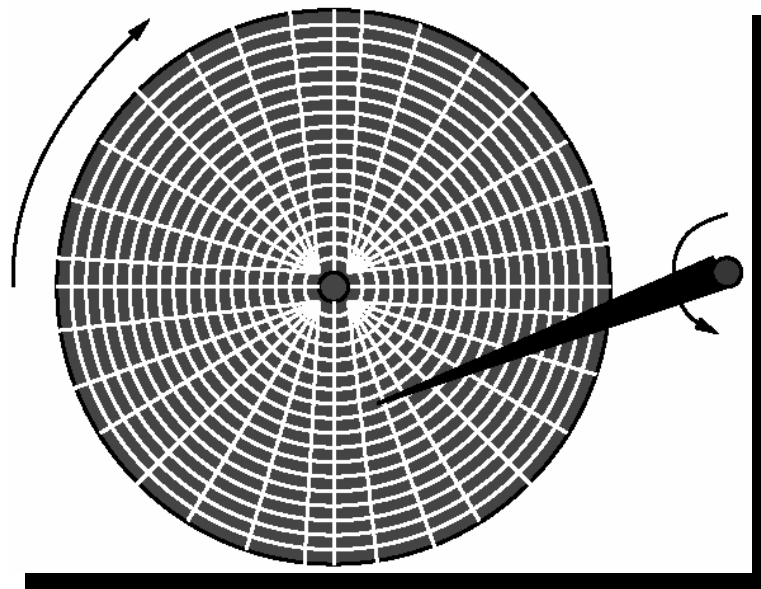
---

- Un disco magnetico consiste di un piatto di alluminio ricoperto da materiale ferromagnetico
- Attualmente le dimensioni del piatto variano da 3 a 12 cm
- La testina del disco scorre sul piatto sfiorando la superficie con un anello posto all'estremità
- In fase di scrittura la corrente che scorre nell'anello magnetizza il materiale sottostante in una direzione che dipende dal verso della corrente
- In fase di scrittura nell'anello viene indotta una corrente con il segno che dipende dal tipo di magnetizzazione indotta in fase di scrittura
- Si ottiene in tal modo un flusso di bit che può essere trasferito nella memoria del calcolatore

# Dischi Magnetici

I dati sulla superficie del disco sono scritti in cerchi concentrici (tracce)

Ogni traccia è divisa in settori che memorizzano un gruppo di dati (512 byte) con le informazioni per correggere eventuali errori e un preambolo per la sincronizzazione



# Dischi Magnetici

---

- 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



# Dischi Magnetici

---

## TEMPO DI ACCESSO

- ***seek time* ( $t_s$ )**: tempo necessario alla testina di lettura per spostarsi da una traccia all'altra.  
Dipende dalle caratteristiche meccaniche del sistema che movimentata la testina
- ***latency time o tempo di latenza* ( $t_L$ )**: tempo necessario per posizionare la testina sul dato da leggere (o scrivere).  
Può valere in media la metà del tempo di rotazione del disco

**Se  $r$  [num. giri al minuto] è la velocità di rotazione il tempo medio di latenza diventa:**

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

# Dischi Magnetici

---

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

**Sono in commercio dischi in cui la velocità di rotazione è da 7000 a 10000 giri al minuto.**

**Sono annunciati dischi fino a 15,5 kgiri al minuto**

**tempo lettura di un dato =  
tempo lettura traccia/numero dati su una traccia =  
=  $(r)^{-1} / N$**

# Caratteristiche dei Dischi

---

## ESEMPI

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

# Tempo di Accesso

---

Tempo di accesso  $t_B$  ad un blocco di lunghezza  $n$ :

$$t_B = t_s + t_L + (\text{tempo di lettura blocco}) =$$

$$t_s + (2r)^{-1} + n (r)^{-1} / N =$$

Nel caso del NEC D2257

$$20 + 8,5 + n * 17 / 20 \text{ (Kbytes) ms} =$$

$$[28,5 + 0,85 * n(\text{Kbytes})] \text{ms}$$

Nel caso del Quantum

$$7,9 + 4,2 + n * 8,4 / 74 \text{ (Kbytes) ms} =$$

$$[12,1 + 0,113 * n(\text{Kbytes})] \text{ms}$$

# Dischi Ottici

---

## CD-ROM:

- Un CD è un disco di policarbonato con una superficie di alluminio riflettente protetta da un film di resina
- Le informazioni sono memorizzate come successioni di "buchi" (pits) e "terre"(lands) nel policarbonato, letti per mezzo di un laser. I dati sono memorizzati a spirale.
- La velocità di scansione è di 1,2m/s costante con una velocità di rotazione che va da 500 a 200 rpm



# Dischi Ottici

---

- I dati sono memorizzati in settori che contengono:
  - Preambolo (16 byte)
  - Dati (2048 byte)
  - ECC, Error Correction Code (288 byte)
- La velocità di lettura è di circa 75 settori/s che corrispondono a circa 150 kbyte/s
- Vi sono diversi modi di memorizzare i dati a seconda che si richieda o meno la presenza di ECC


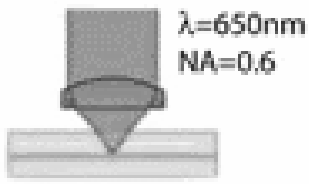


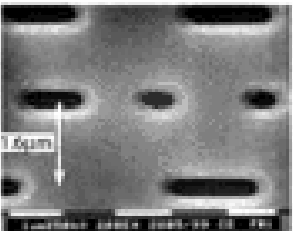
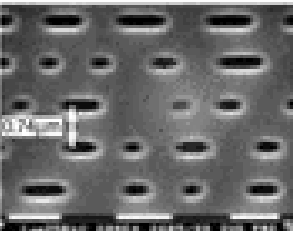
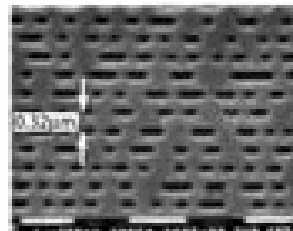
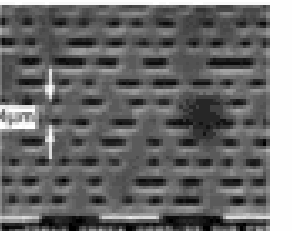
# Dischi Ottici Riscrivibili

---

## ■ CD-RW:

- Strato alluminio sostituito da lega con due stati stabili, con diverse capacità riflettenti:
  - cristallino (riflettente)
  - amorfo (semi riflettente)
- Laser a tre potenze:
  - alta potenza porta allo stato amorfo
  - media potenza porta allo stato cristallino
  - bassa potenza lettura

# Dischi Ottici

| CD  | DVD   | BD  | HD-DVD   |
|---|---|---|--|
|  <p><math>\lambda=780\text{nm}</math><br/><math>NA=0.45</math></p> <p>2.1 <math>\mu\text{m}</math></p> |  <p><math>\lambda=650\text{nm}</math><br/><math>NA=0.6</math></p> <p>1.3 <math>\mu\text{m}</math></p> |  <p><math>\lambda=405\text{nm}</math><br/><math>NA=0.85</math></p> <p>0.6 <math>\mu\text{m}</math></p> |  <p><math>\lambda=405\text{nm}</math><br/><math>NA=0.65</math></p> <p>0.76 <math>\mu\text{m}</math></p> |
|  <p>1.6 <math>\mu\text{m}</math></p>   |  <p>0.9 <math>\mu\text{m}</math></p>  |  <p>0.35 <math>\mu\text{m}</math></p>  |  <p>0.4 <math>\mu\text{m}</math></p>  |

CD: dati e eaudio => 650 MB

DVD: dati e video => 4,7 GB

BD: Blu-ray Disc® HDTV => 25-200 GB

HD-DVD: HDTV => 25 GB

DVD strato singolo - singola facciata (4,7 GB)

DVD strato doppio - singola facciata (8,4 GB)

DVD strato singolo - doppia facciata (9,4 GB)

DVD stato doppio - doppia facciata (17 GB)

# Il Sistema Memoria

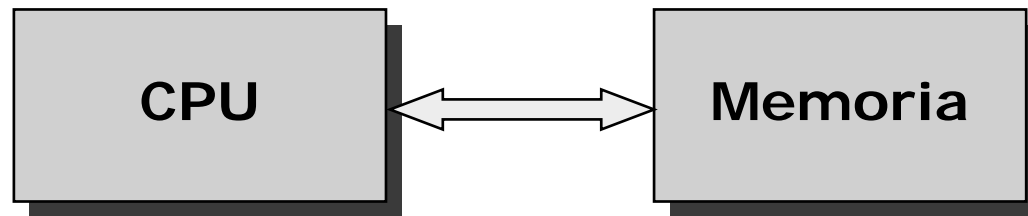
---

- 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 architettonico è quello di raggiungere un ragionevole compromesso fra gli obiettivi indicati

# CPU-Memoria

---

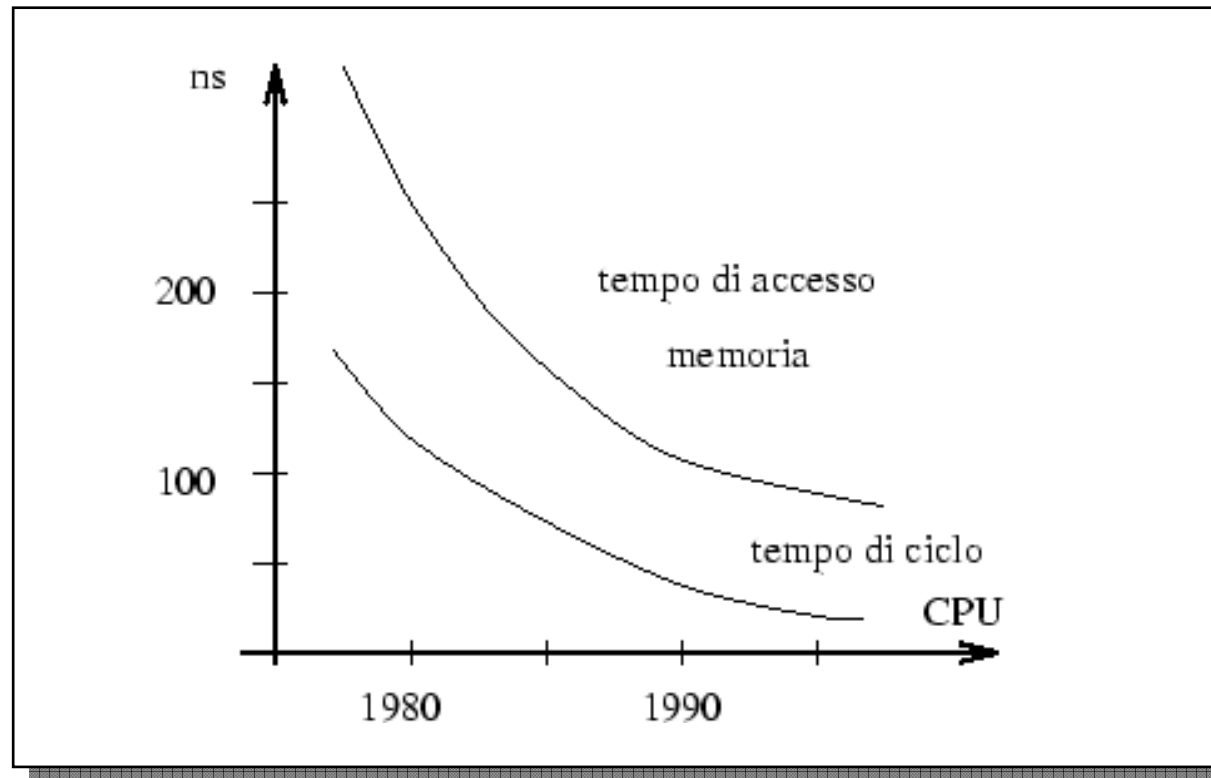
Nell'architettura VonNeuman il canale di comunicazione tra la CPU e la memoria è il punto critico (collo di bottiglia) del sistema



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

# CPU-Memoria

---



Sono disponibili nel 2006 CPU con frequenza di clock superiore a 3 GHz  
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



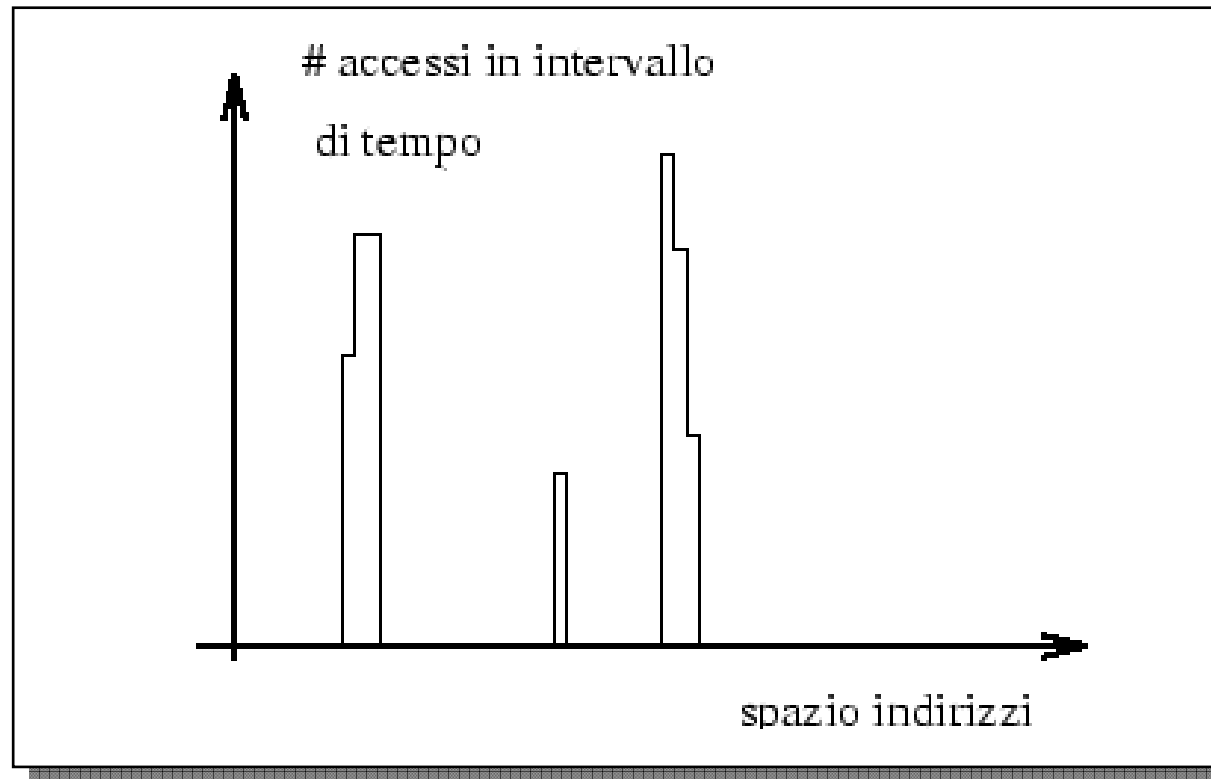
# Principio di Località

---

- 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  $X + \Delta X$  sia richiesto fra  $t$  e  $t + \Delta t$ "
- Nel breve periodo gli indirizzi generati da un programma sono confinati in regioni limitate

# Principio di Località

---



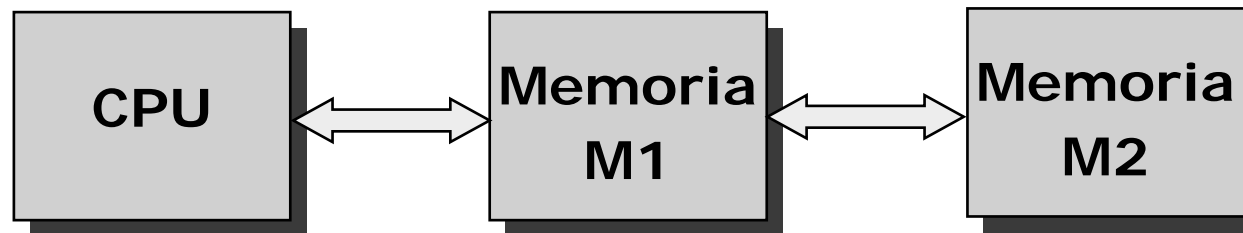
La distribuzione degli accessi alla memoria in un dato intervallo può essere misurato direttamente su un sistema

# La Gerarchia delle Memorie

---

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
  
- 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 di Accesso

---

Tempo di accesso medio globale:

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

dove

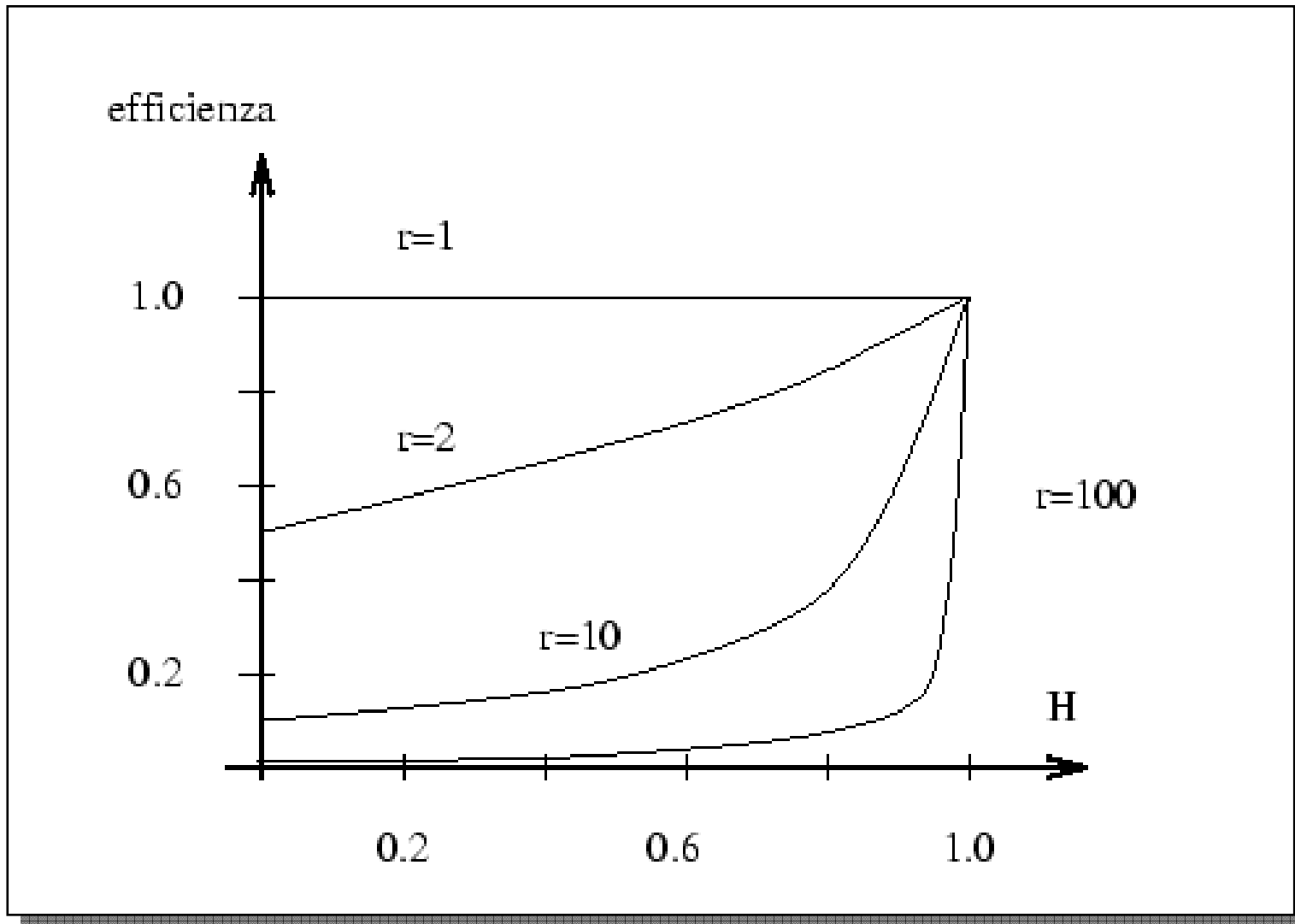
- $t_{A2} = t_{A1} + t_B = r t_{A1}$
- $t_B$  è il tempo di accesso a un blocco di  $M_2$

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:  $r = \frac{t_{A2}}{t_{A1}}$

# Prestazioni



# Memoria Virtuale

---

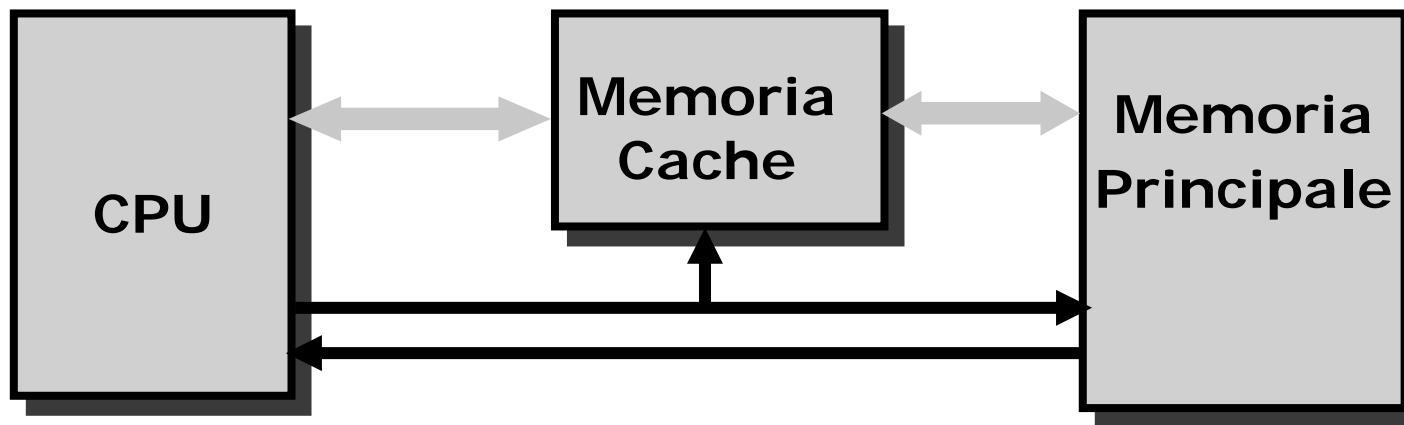
- La Memoria Virtuale è il meccanismo di gestione della gerarchia formata da:
  - Memoria principale
  - Memoria secondaria
  
- Offre il grosso vantaggio di rendere il programma indipendente dalle dimensioni di memoria della singola macchina



# Memoria Cache

---

- La cache è una memoria veloce e di piccole dimensioni posta fra la CPU e la memoria principale
- La cache e la memoria principale formano una gerarchia di memoria



I tre livelli di memoria sono gestiti come due gerarchie:

- Mem. principale - Mem. secondaria
- Cache - Memoria principale

# Memoria Cache

---

- Le prestazioni della memoria cache dipendono anche dalla sua posizione rispetto alla CPU:
  - Cache su scheda
  - Cache su chip
- La presenza di una memoria cache sullo stesso chip del processore rappresenta la soluzione che garantisce la maggiore efficienza
- Nei sistemi più recenti sono presenti entrambe le soluzioni

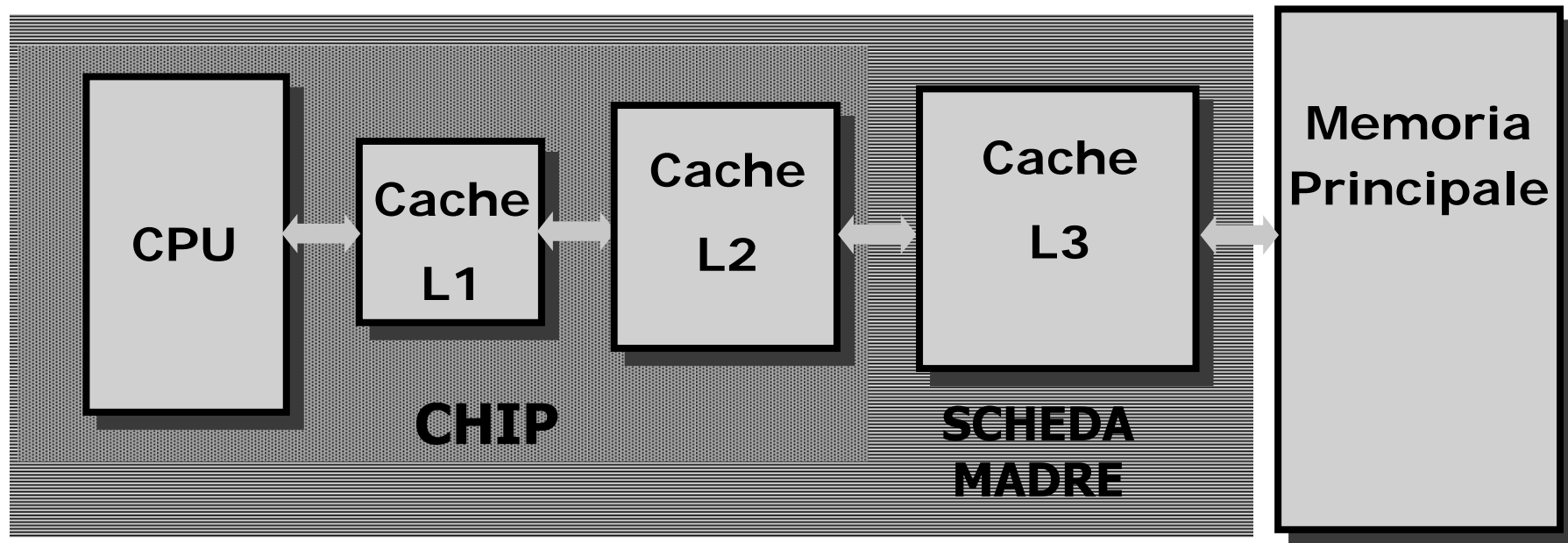
# Prestazioni a confronto

---

| Gerarchia                 | Cache - M.Pr. | M.Pr. - M.Sec.          |
|---------------------------|---------------|-------------------------|
| Rapporto tempi di accesso | 5/1           | 1000/1                  |
| Gestione                  | Hardware      | Software                |
| Dim. pagina               | 4 - 128 bytes | 64 - 4096 bytes         |
| Accesso CPU al 2o livello | Diretto       | Sempre tramite il primo |

# Memoria Cache

Nei processori moderni esistono fino a 3 livelli di cache (L1, L2, L3)



**Soluzione L1+L2**

**L1 ≥ 32 kB**

**L2 ≥ qualche MB**

**Soluzione L1+L2 + L3**

**L1 ≥ 16 kB**

**L2 ≥ 256 kB**

**L3 ≥ 2-4 MB**