

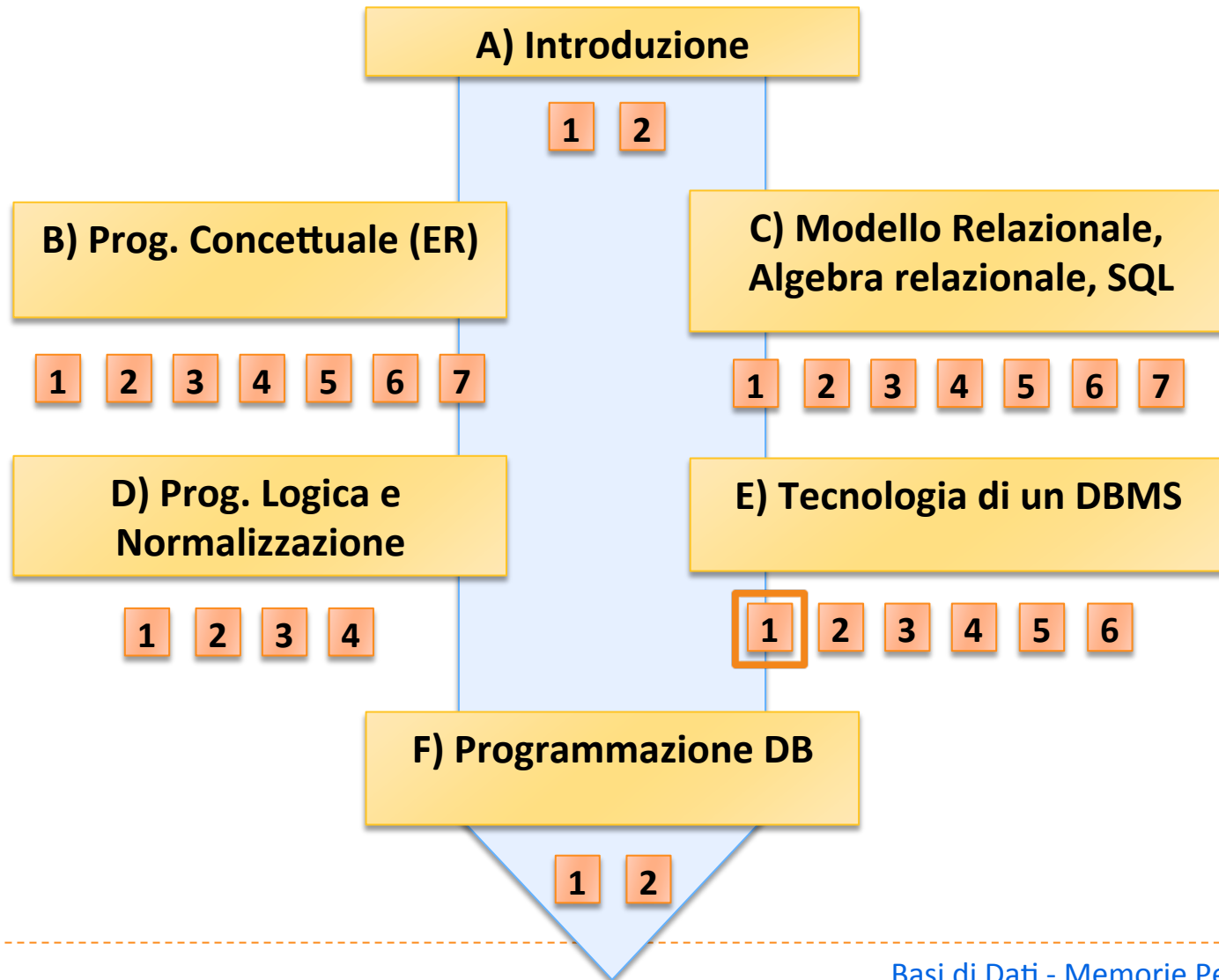


**Basi di Dati**



**Memorie Permanenti**

# Basi di Dati – Dove ci troviamo?



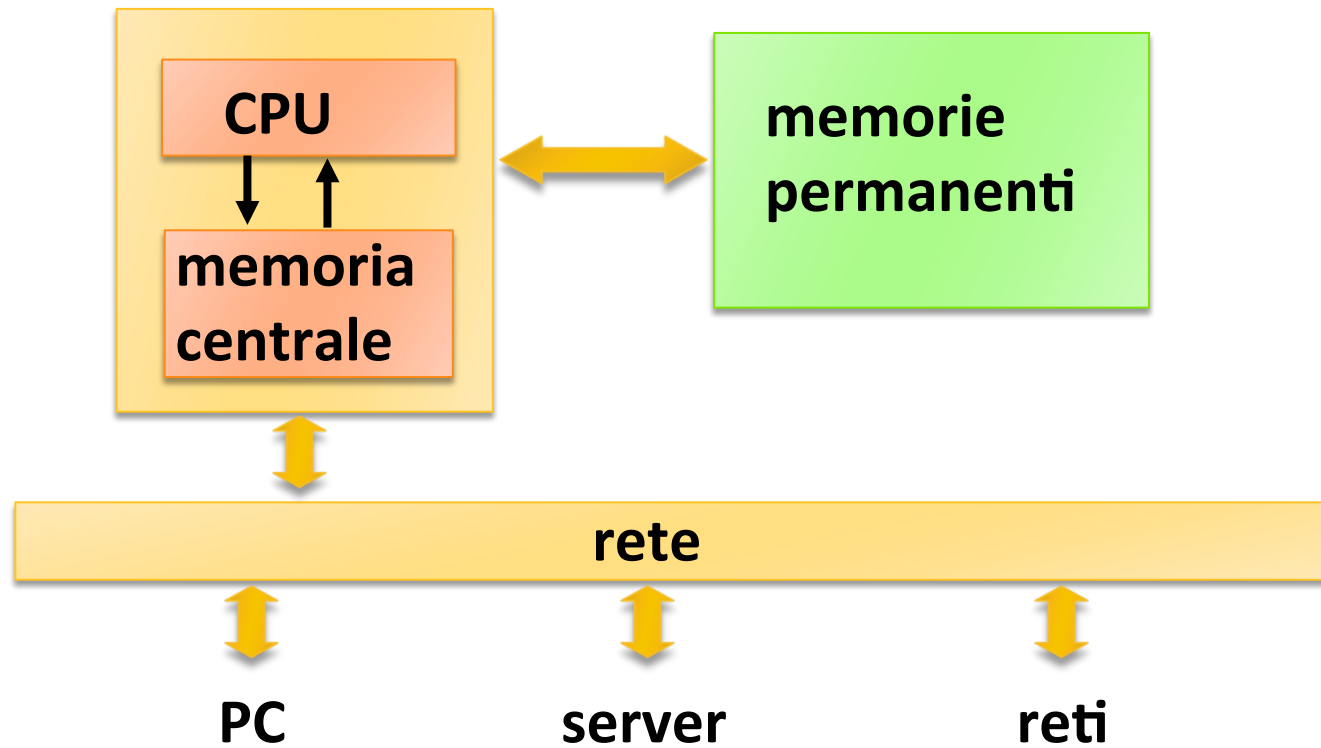
# In questa lezione

---

- ▶ **Presenteremo:**
  - ▶ le unità di **memoria permanente**
  - ▶ il loro funzionamento
  - ▶ le caratteristiche principali
  - ▶ l'affidabilità

# Struttura di un data server

---



# Qualità di un data server

---

- ▶ velocità della **CPU**
- ▶ capacità e velocità della **memoria centrale** (...o memoria di servizio...)
- ▶ capacità e velocità delle **memorie permanenti** (...o memorie secondarie...)

si tende ad enfatizzare le prime due mentre la più importante è la terza!

perché condiziona la velocità del servizio nelle **applicazioni gestionali**

# Qualità di un data server

---

- ▶ La **qualità della CPU**, a parità di tecnologia elettronica costruttiva, si misura in numero di **Mhz** del clock e in **numero di bit** dei registri (32-64)
- ▶ le **prestazioni** generali tendono a migliorare di circa 1.5 volte ogni anno (negli ultimi anni)
- ▶ i **costi** sono in forte calo ( a parità di prestazioni)

# Qualità di un data server

---

- ▶ secondo la **legge di Amdahl**, tenendo ferma la tecnologia delle memorie permanenti ed aumentando la velocità della **CPU** di un fattore 10, si avrebbe un miglioramento delle prestazioni del server di un fattore 5; aumentando di un fattore 100 si avrebbe un miglioramento di solamente 10
- ▶ ciò giustifica lo sforzo dell'industria per **adeguare le prestazioni** delle memorie permanenti

## Legge di Amdahl (dei rendimenti decrescenti):

il miglioramento di una delle componenti di una macchina non produce un aumento delle prestazioni proporzionale al miglioramento. Il possibile incremento delle prestazioni è limitato dall'ammontare dell'utilizzo del componente.

# UNITA' STANDARD

---

grandezza	nome	abbreviazione
▶ $10^{15} \div 2^{50}$	peta	p, P
▶ $10^{12} \div 2^{40}$	tera	t, T
▶ $10^9 \div 2^{30}$	giga	g, G
▶ $10^6 \div 2^{20}$	mega	m, M
▶ $10^3 \div 2^{10}$	kilo	k, K
▶ $10^{-3}$	milli	m
▶ $10^{-6}$	micro	$\mu$
▶ $10^{-9}$	nano	n
▶ $10^{-12}$	pico	p



# UNITA' STANDARD

---

▶ hertz (1 ciclo al sec.)	Hz,hz
▶ bit	b
▶ byte (8 bit)	B
▶ bits(bytes) per second	bps (Bps)
▶ instructions per sec.	ips
▶ I/O operations per sec.	I/Ops
▶ transactions per sec.	tps
▶ bits per inch (2.5 cm)	bpi
▶ rounds per minute	rpm

# Capacità presente e futura dei sistemi di memoria

---

**Presenti:**

**alcuni Giga( $10^9$ )**

**Terabytes ( $10^{12}$ )**

**Petabytes ( $10^{15}$ )**

**vicine:**

**Exabytes ( $10^{18}$ )**

**Zettabytes ( $10^{21}$ )**

**future:**

**Yottabytes( $10^{24}$ )**

# Un'idea sull'hardware

---

**PC di oggi: 3 Ghz multi-core, 4096 MB RAM,  
2 TB di disco, velocita' massima: fino a 100 Gigaflops**

**Tianhe-1 (NSC, 2010): 2,507 Petaflops, 14336 processori,  
262 Terabyte RAM, 2 PetaByte di disco (88M \$ + 20M \$/anno)**

**Cray Jaguar (Cray, 2009): 1,75 Petaflops, 224256 processori,  
360 Terabyte RAM, 10 PetaByte di disco (104M \$)**

**IBM RoadRunner (IBM, 2008): 1,7 Petaflops, 12960+6480  
processori, 103,6 Terabyte RAM, (133M \$)**

# Ordini di grandezza

---

## TESTI:

- ▶ **1 byte: 1 carattere**
- ▶ **1 pag. di libro : 50 righe per 80 caratteri → 4 kB**
- ▶ **1 libro di 500 pag.: 2MB (senza figure)**
- ▶ **1 pag. di vocabolario: 2×60×80 caratteri → 9.6 kB**
- ▶ **1 vocabolario di circa 2000 pag. → 20 MB**
- ▶ **att.: l'occupazione di memoria è superiore se le pagine ed i caratteri sono "strutturati"**
- ▶ **in 500MB di 1 CD vanno 250 libri o 25 vocabolari**
- ▶ **in un BD-ROM da 50 GB: 25000 libri o 2500 vocabolari**
- ▶ **Nota: con tecniche di compressione è possibile incrementare anche di molto queste cifre**

# Ordini di grandezza

---

## IMMAGINI:

- ▶ Immagini di 1000×1000 **pixel** a seconda dei livelli di grigio o dei livelli dei tre colori base : da 1 a 4MB
- ▶ Immagini ad **alta definizione** (pixel di 25 μm di lato): 100MB
- ▶ Le immagini possono essere **compresse**.
- ▶ Da una foto aerea si può ottenere una carta come immagine (immagine "**raster**") e successivamente una rappresentazione "**vettoriale**" cioè per linee e punti riducendola a circa 100kB.

# Tipi di memorie permanenti

---

- ▶ **memorie elettroniche**
  - ▶ **memorie flash**
- ▶ **memorie magnetiche**
  - ▶ **dischi rigidi**
    - ▶ disco singolo
    - ▶ RAID (dischi paralleli)
- ▶ **memorie ottiche**
  - ▶ **CD-ROM, CD-R, CD-RW**
  - ▶ **DVD-ROM, DVD-R, DVD-RW**
  - ▶ **BD-ROM, HD-DVD**
  - ▶ **magneto-ottiche riscrivibili**

# Memorie elettroniche - flash

---

- ▶ **Esistono vari tipi di memorie elettroniche permanenti. Molto comuni sono le **Flash Memory**, usate per memorizzare informazioni in modo veloce e semplice, come dei piccoli hard disk.**
- **Alcuni esempi di Flash Memory:**
  - Il BIOS del computer
  - “Chiavi” USB
  - Schede CompactFlash, Memory Stick, SD (fotocamere digitali)
  - Dischi SSD (Solid State Drive)
  - Memory card usate nelle console



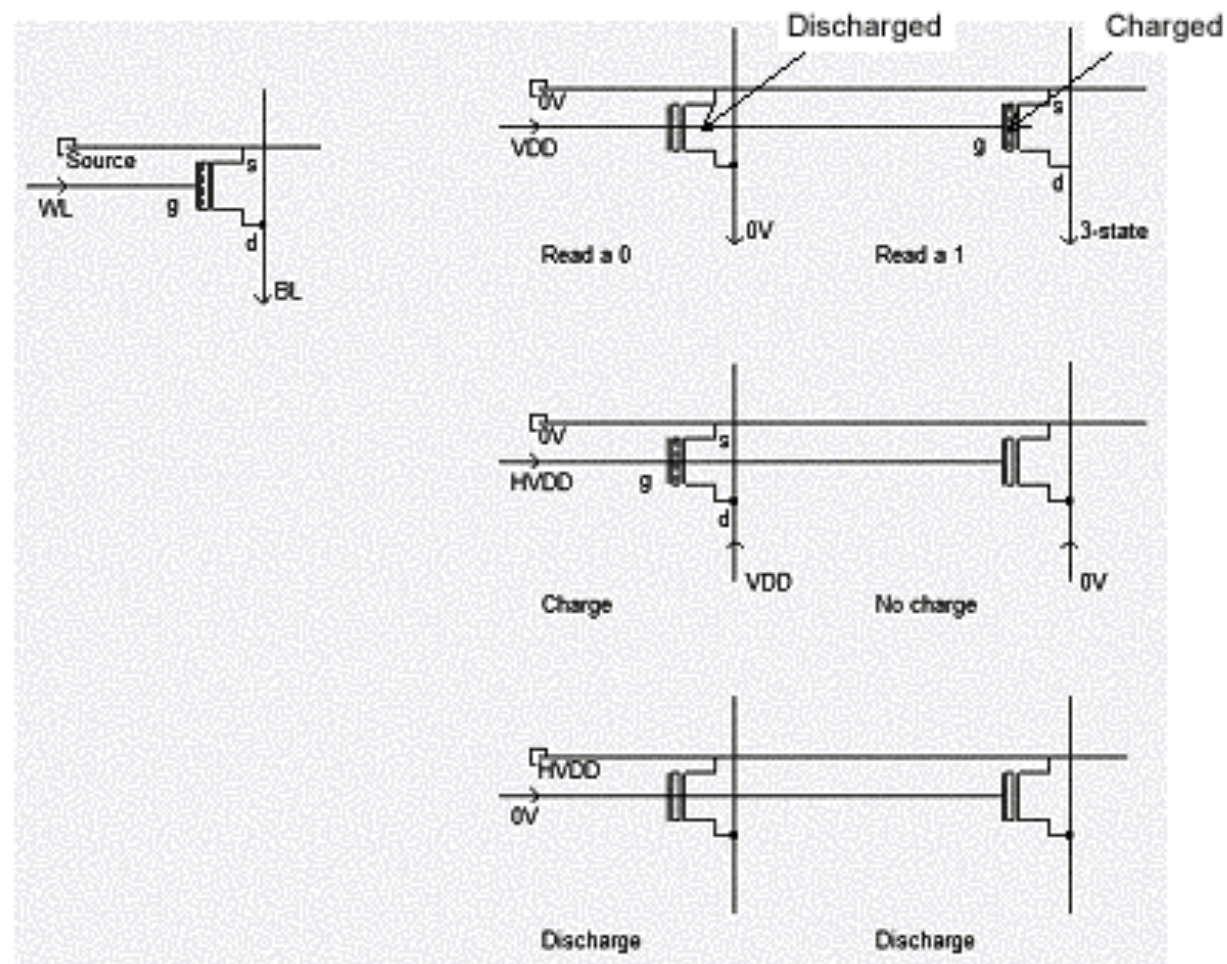
# Memorie flash

---

- ▶ Le memorie Flash sono ROM di tipo EEPROM che hanno una griglia formata da righe (word line WL) e colonne (bit line BL). Le celle ad ogni intersezione hanno un **transistor** con doppio gate. I gate sono separati da un finissimo strato di ossido. Tra i due gate può venire immagazzinata della carica, che determina il valore della cella.
- ▶ Uno dei gate viene definito **Control Gate** mentre l'altro **Floating Gate**. Il Floating Gate è collegato alle righe attraverso il Control Gate.



# Memorie flash



# Memorie flash

---

## ▶ **Carica:**

- ▶ Operazione selettiva (dipende dall'informazione sulla bit line)
- ▶ Sovratensione HVDD applicata sui gate
- ▶ Tensione VDD sulla bit line: iniezione di carica (scrivo 1)
- ▶ Tensione 0 sulla bit line: nessuna iniezione di carica (mantengo 0)

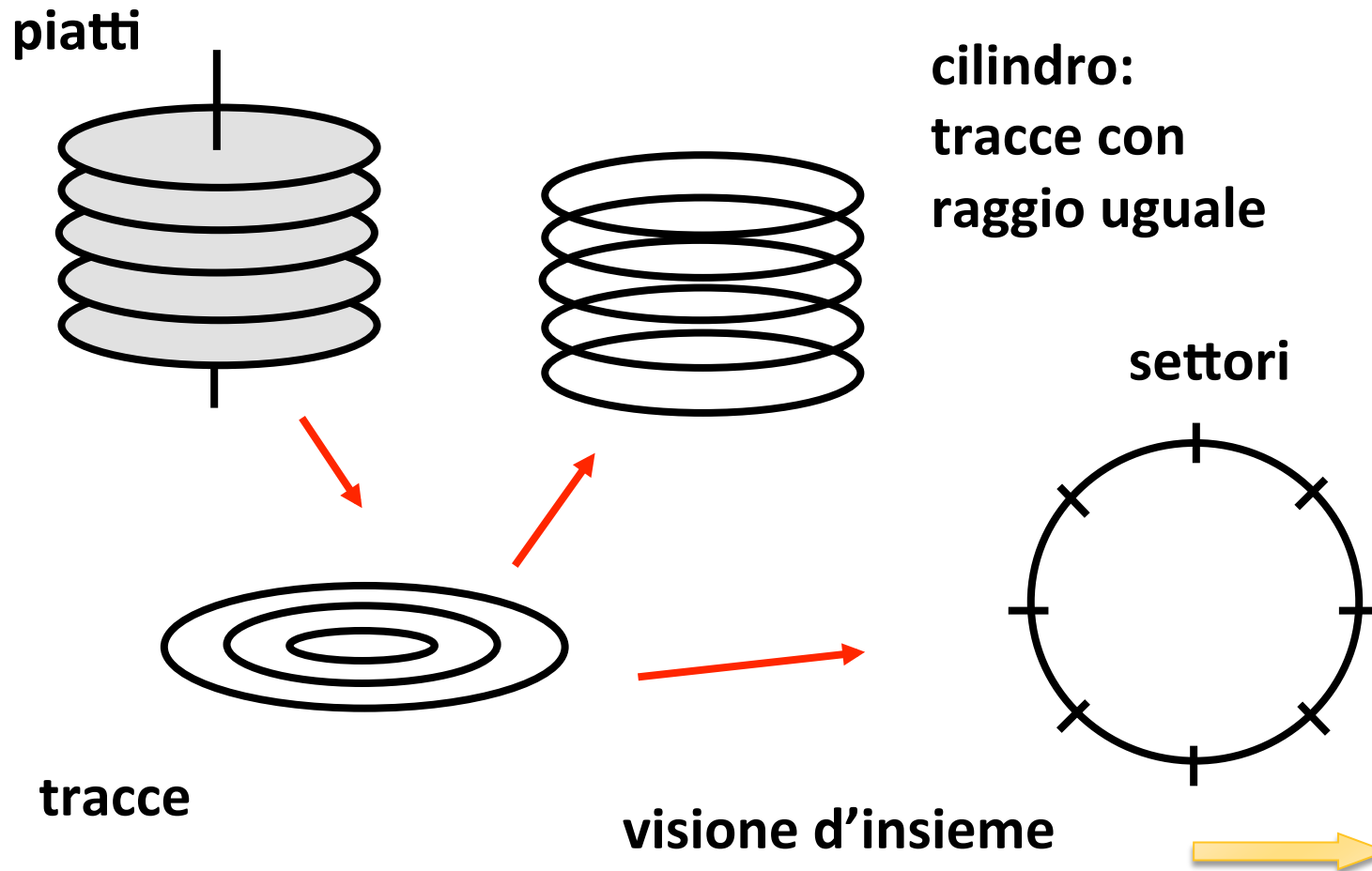
## ▶ **Scarica:**

- ▶ Operazione non selettiva (si scaricano tutte le celle di una WL)
- ▶ Sovratensione HVDD applicata sui source
- ▶ Tensione nulla sui gate

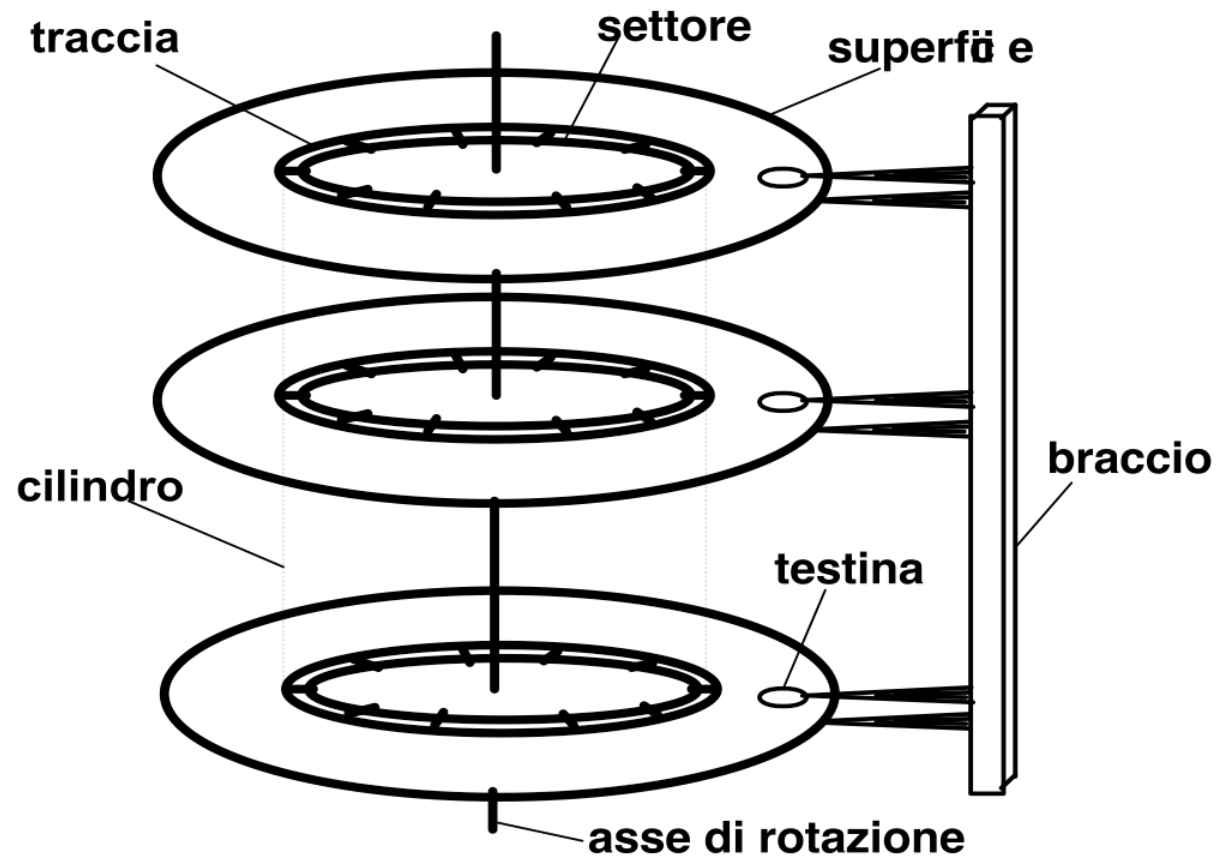
## ▶ **Lettura:**

- ▶ Si applica una tensione di alimentazione VDD al control gate e si mette a massa la WL
- ▶ Sulla bit line si legge 0 / 1

# Memorie magnetiche – il disco



# Memorie magnetiche – il disco



# Grandezze

---

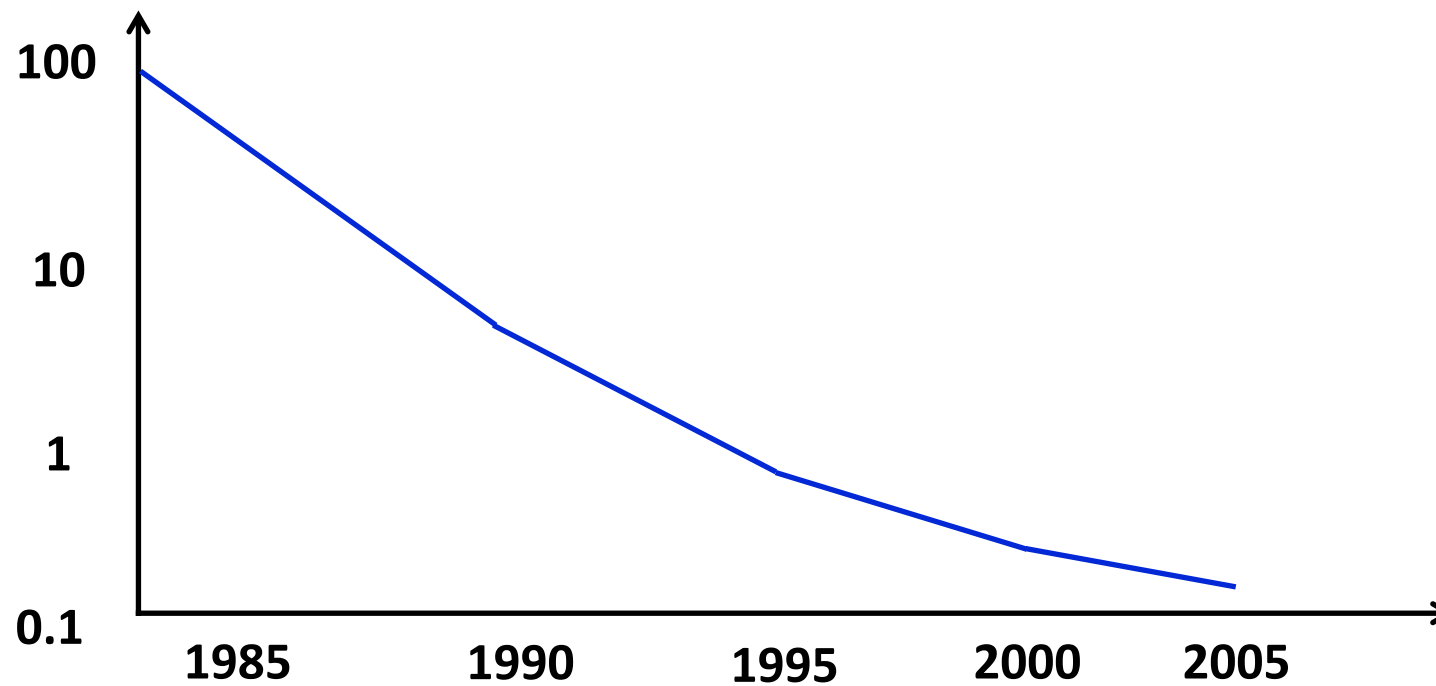
- ▶ **velocità di rotazione (rpm):** 5400 → 10000 ed oltre
- ▶ **no. piatti:** 1 → 20 ed oltre
- ▶ **tracce per piatto (quindi cilindri)** 3000 → 20000 ed oltre
- ▶ **diametro disco** 1 → 8 inch
- ▶ **densità:** n Giga bit per inch<sup>2</sup>
- ▶ **dimensione del settore:** 4k ed oltre
- ▶ **settori per traccia:** ( $2^n$  ad es. 64) da circa alcune decine ed oltre



# Grandezze

---

Costo in \$ per MB (pc)



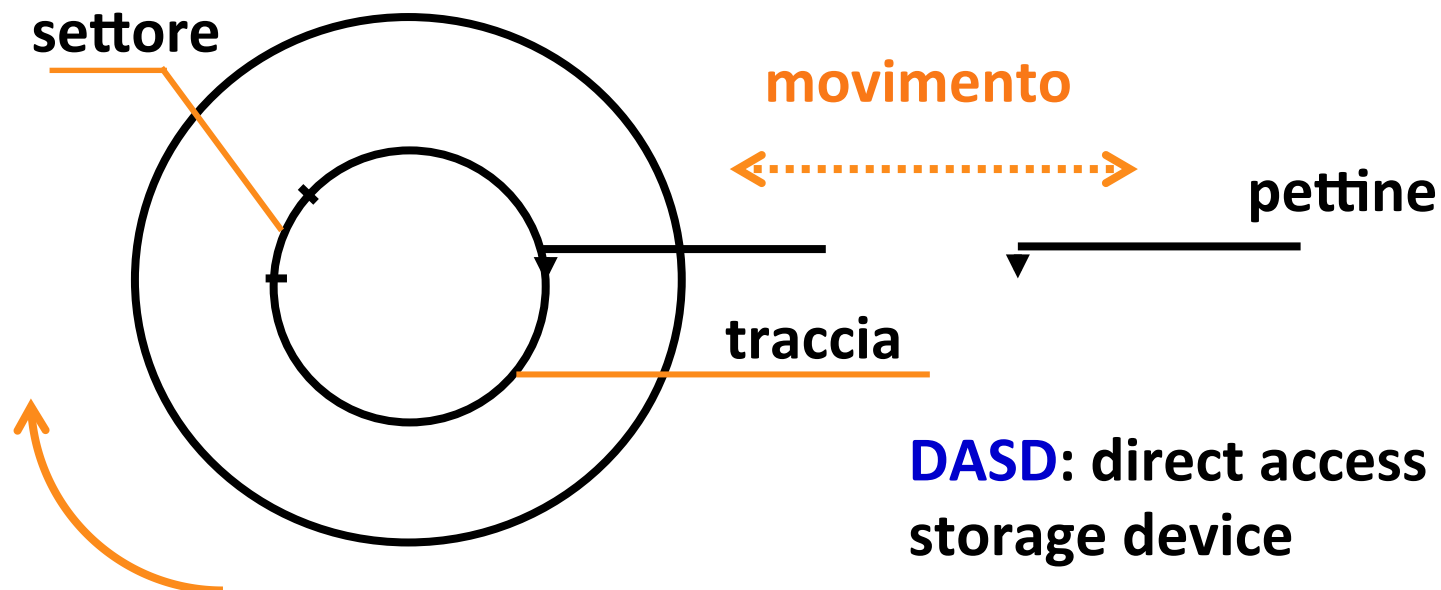
# Dimensione del settore

---

- ▶ Ad esempio, con i file system in **Windows**:
  - ▶ **NTFS** (Windows NT, 2000, XP, 2003, Vista...)
    - ▶ 4K block size per dischi più grandi di 2GB
  - ▶ **FAT-32** (Windows 95, 98, ME)
    - ▶ 4K per dischi fino a 8GB
    - ▶ 8K fino a 16GB
    - ▶ 16K fino a 32GB
    - ▶ 32K oltre 32GB.

# Meccanica del disco

Funzionamento: movimento del **pettine**,  
raggiungimento del **cilindro** richiesto,  
attivazione della testa relativa alla **traccia**,  
attesa del **settore** , lettura/scrittura





# Meccanica del disco

---

- ▶ Il **settore** è l'unità minima di trasferimento, i settori possono essere raggruppati in **blocchi** (pagine)
- ▶ l'**indirizzo** di un settore (blocco) è :
  - ▶ **num. cilindro, num. traccia, num. settore.**
- ▶ il **tempo di servizio** è:
  - ▶ tempo di **posizionamento** (seek time) :  **$T_s$**
  - ▶ tempo di **latenza rotazionale** :  **$T_r$**
  - ▶ tempo di **lettura** (scrittura) :  **$T_b$**
  - ▶ per la scrittura si usa anche il metodo **read after write** che ricontrolla dopo un giro
  - ▶ tempo impiegato dal **controller** (elett.):  **$T_c$**

# Meccanica del disco

---

- ▶ il tempo di posizionamento (**seek time**) : **T<sub>s</sub>** viene indicato dal costruttore come **tempo medio di spostamento** tra due possibili tracce, vengono anche indicati il **T<sub>max</sub>** ed il **T<sub>min</sub>**.
- ▶ il tempo di **latenza rotazionale** : **T<sub>r</sub>** è mediamente la metà del tempo di rotazione
- ▶ il tempo di **lettura** (scrittura) : **T<sub>b</sub>** dipende dalla dimensione del blocco
- ▶ il metodo **read after write** richiede un ulteriore  $2 \times T_r$
- ▶ il tempo impiegato dal **controller** (elett.): **T<sub>c</sub>** è generalmente indicato dal costruttore
- ▶ **transfer rate** misurato in MB/sec.

# Meccanica del disco

---

▶ **Esempio:**

**$T_s = 9\text{ms}$ , transfer rate = 30 MB/sec ,**

**blocco = 4096 bytes,  $T_c = 1\text{ ms}$ .**

**rotazione 7200 rpm**

▶ **tempo di accesso:**

$$T_s + T_r + T_b + T_c =$$

$$9\text{ ms} + 0.5 / 7200\text{rpm} + 4\text{KB} / 30\text{MB/sec} + 1$$
$$\text{ms} = 9 + 4.15 + 0.1 + 1 = \mathbf{14.3\text{ ms}}$$

▶ **con read after write :**

$$14.3 + 2 \times 4.15 = \mathbf{22.6\text{ ms}}$$

**l'ordine di grandezza è di molto superiore a quello delle operazioni elettroniche**

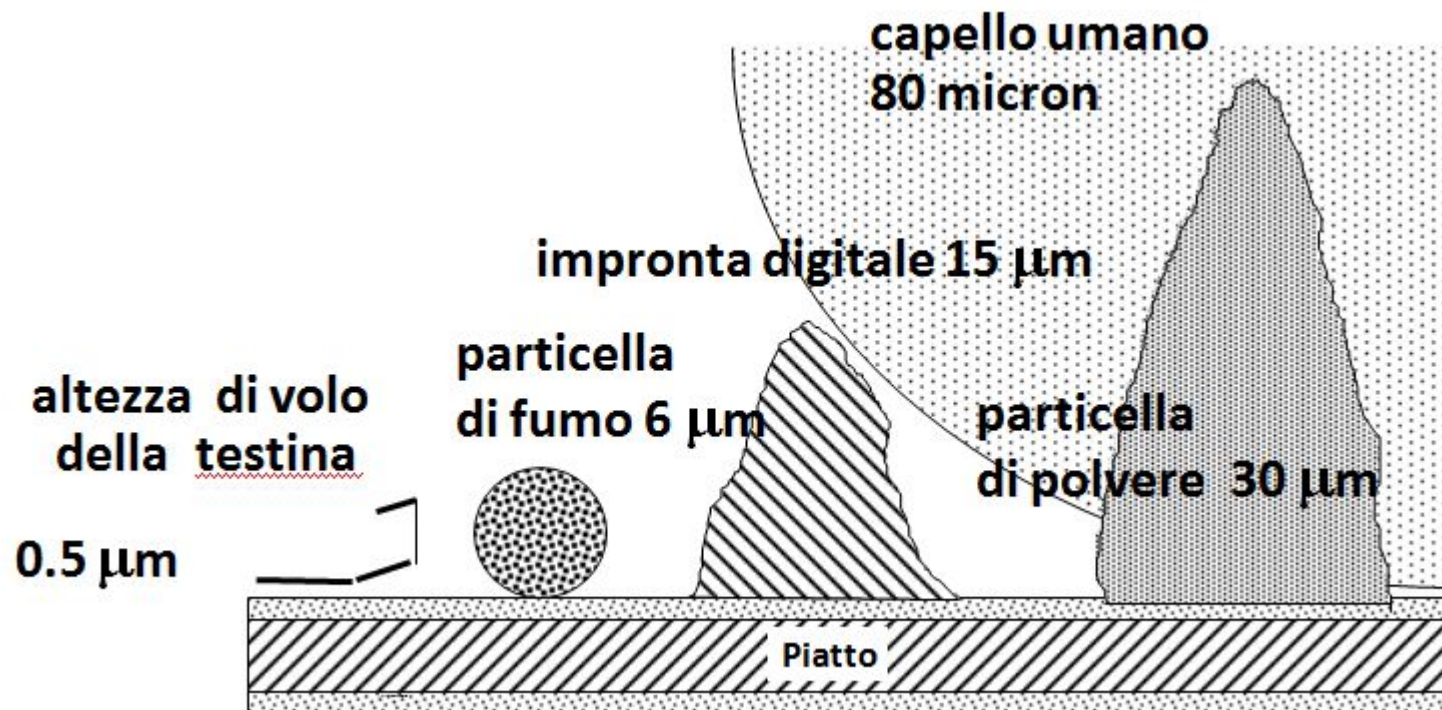
# Meccanica del disco

---

- ▶ Le **memorie elettroniche** avrebbero un tempo di accesso di circa 100000 volte inferiore ... ma un costo di circa 100 volte superiore
  - ▶ i tempi di **seek** e di **latenza** sono da ridurre:
    - costruttivamente, riducendo i **diametri**, aumentando la **velocità** del pettine e di rotazione, aumentando la **densità dei bit** sulla superficie e riducendo di conseguenza **l'altezza di volo** delle testine sui piatti.....
- l'altezza** di volo è inferiore a **0.5 micron** (si pensi che un capello ha un diametro di circa **80**, la polvere da **6 a 30** ed un'impronta digitale di **15 micron**)

# Meccanica del disco

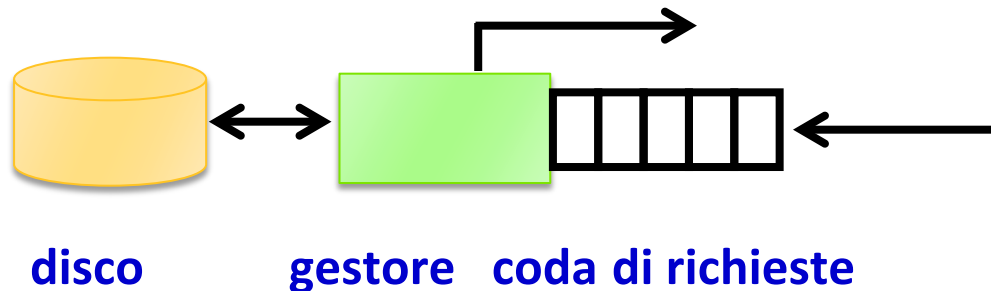
L'**altezza di volo della testina** è strettamente connessa alla densità di registrazione: minore è la distanza, migliore è la possibilità di rilevare le variazioni di campo magnetico, maggiore è la densità possibile.



# Meccanica del disco

---

- riducendo l'ampiezza degli **spostamenti** del pettine organizzando la **coda** delle richieste dell'utenza:



I **criteri** più noti per soddisfare le richieste sono i seguenti ( o varianti e combinazioni degli stessi):



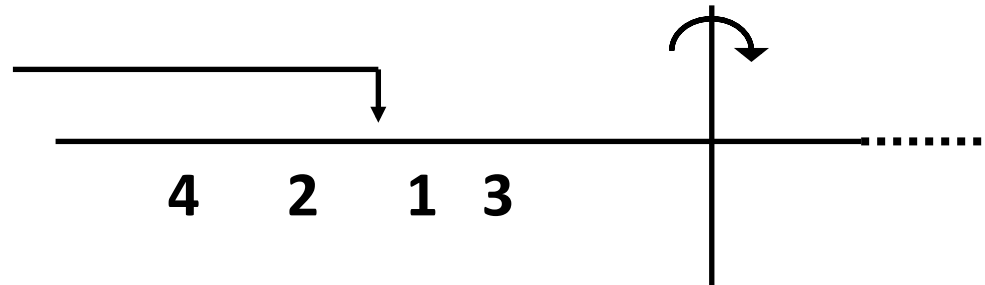
# Meccanica del disco

---

- selezione dipendente dal **richiedente**:
  - FIFO** : First In First Out
  - PRI** : priorit  dipendente dal processo
  - PRI/FIFO** : combinazione delle due;
- selezione dipendente dall'**oggetto** richiesto:
  - SSTF** : Shortest Seek Time First,
  - SCAN**: SSTF in una sola direzione, in avanti e poi indietro sul disco,
  - C-SCAN**: SSTF in avanti con ritorno veloce,
  - N-STEP\_SCAN**: SCAN di solo N settori per volta (tecnica di Disk Sharing).

# Meccanica del disco

---



- FIFO** : 1, 2, 3, 4    imparziale ma lenta
- SSTF** : 2, 4, 1, 3    sfavorisce le richieste  
   lontane dalla testa
- SCAN** : 1, 3, 2, 4    buona
- C-SCAN**: 1, 3, 4, 2    buona



# Memorie ottiche

---

- ▶ La più nota è il **CD-ROM**:
- ▶ funzionamento **start-stop**
- ▶ **accesso diretto** ( $n \times 100$  ms  
con data rate  $n \times 100$  KB/sec)
- ▶ **molta** memoria (600 MB  $\rightarrow$  oltre)
- ▶ unità di **sola lettura**
- ▶ **trasporto** archivi medio/grandi
- ▶ costituiscono una memoria di terzo livello
- ▶ cabinet di tipo **juke-box** con memoria da  $n \times 100$   
GB  $\rightarrow$   $n \times$  TB



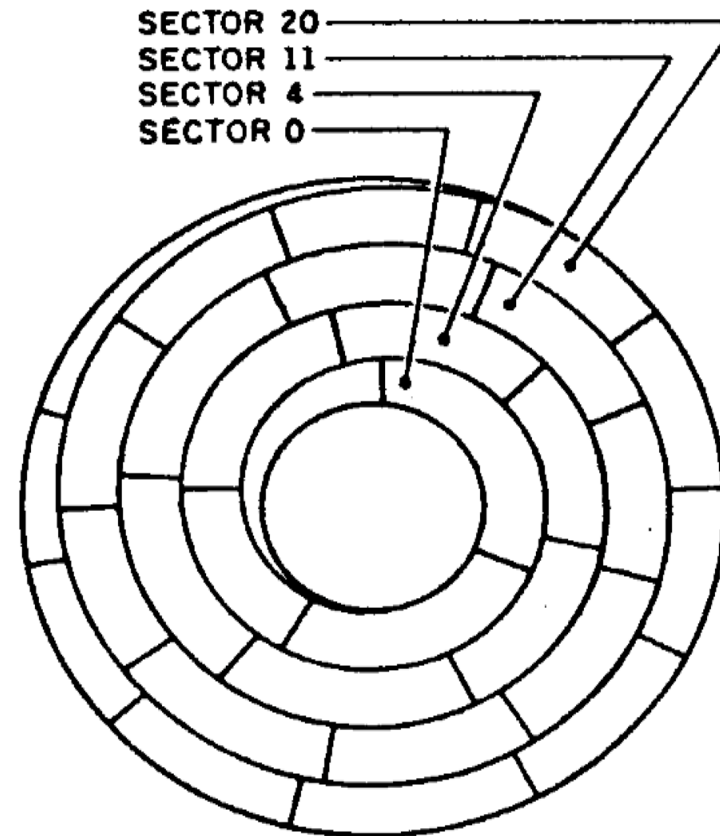
# Memorie ottiche

dischi magnetici  
sono **CAV**: constant  
angular velocity

i CD-ROM sono **CLV**:  
constant linear velocity

un laser legge su una  
superficie riflettente  
la presenza di fori  
( $\varnothing < 1\mu\text{m}$ ) provocati  
da un laser di potenza

unica traccia a **spirale**  
(n×km per un diam. di 12cm)



# Memorie ottiche

---

- ▶ Simile al CD-ROM è il **DVD-ROM**
- ▶ Tecnologia di lettura / scrittura simile a quella dei CD
- ▶ Maggiore quantità di dati contenibile, grazie a...
  - ▶ **alta densità** dei dati (distanza tra le spire di 740 nanometri, contro i 1600 dei CD)
  - ▶ tecnologia **multi-layer**, che consente di raggiungere 15.9GB per DVD doppio lato e doppio strato
  - ▶ Migliori algoritmi di **error correction**, che consentono un minore spazio sprecato da informazioni duplicate

# Memorie ottiche

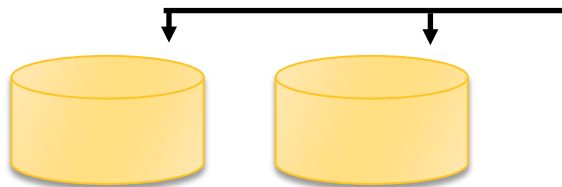
---

- ▶ L'ultima generazione di memorie ottiche prevede:
  - ▶ **HD-DVD**
  - ▶ **Blu-Ray Disk - BD**
- ▶ **HD-DVD**
  - ▶ laser a **luce blu (405 nm)** per una maggiore densità dei dati
  - ▶ tecnologia **multi-layer**, che consente di raggiungere 60GB per HD-DVD doppio strato e doppio lato
- ▶ **Blu-Ray Disk**
  - ▶ laser a **luce blu (405 nm)** per una maggiore densità dei dati
  - ▶ tecnologia **multi-layer**, che consente di raggiungere 50GB per BD doppio strato (BD-50)

# Parallelismo e sicurezza

---

**Disk Mirroring** : dischi con dati replicati



letture indipendenti  
scritture su entrambi

- aumentando il numero di dischi aumenta la probabilità di averne uno **guasto**
- diminuisce la probabilità di **perdita** dei dati dovuta al guasto contemporaneo
- **ridondanza** eccessiva

**la ridondanza è utile!**

# Dischi RAID

---

## Redundant Array of Inexpensive Disks



- architettura che migliora le **prestazioni** e l'**affidabilità** del sistema di mem. permanente
- l'uso di **N** dischi consente di suddividere i dati in **piccoli blocchi** da scrivere e leggere in **parallelo**
- informazioni ridondanti consentono la **correzione** di errori dovuti a guasti

# Dischi RAID

---



- **servizio parallelo** indipendente per **letture brevi** per più utenti (parallelismo inter-query)
- **servizio parallelo** per **letture lunghe** per lo stesso utente (parallelismo intra-query)
- **RAID 0** : striping, nessuna ridondanza (es. 8 dischi)
- **RAID 1** : mirroring (es. 16 dischi)
- **RAID 2-3**: striping, unico movimento parallelo delle testine
- **RAID 4**: striping, prevede un disco in più per l'informazione ridondante : **la parità**

# Dischi RAID

---

## Concetto di parità:

- dati 8 bit c'è un nono bit che contiene la somma **modulo 2** (è 0 se il numero di 1 nel byte è pari altrimenti è 1)

0 0 1 0 1 1 0 1 → 0

1 0 1 1 0 1 1 0 → 1

- se un bit **si inverte** la parità **non torna** quindi c'è un bit errato (però non si sa quale)
- nel **RAID 4** **tutti bit** di parità stanno sul **nono** disco



# Dischi RAID

---



- Se un disco si **guasta** il **controller** se ne accorge e l'informazione viene ricostruita con la **parità**:

**0 0 1 0** x **1 0 1** → **0** poiché la parità è **0** il bit era **1**

**1 0** x **1 0 1 1 0** → **1** poiché la parità è **1** il bit era **1**

Il **RAID 4** ha un **eccesso** di letture sul nono disco

# Dischi RAID

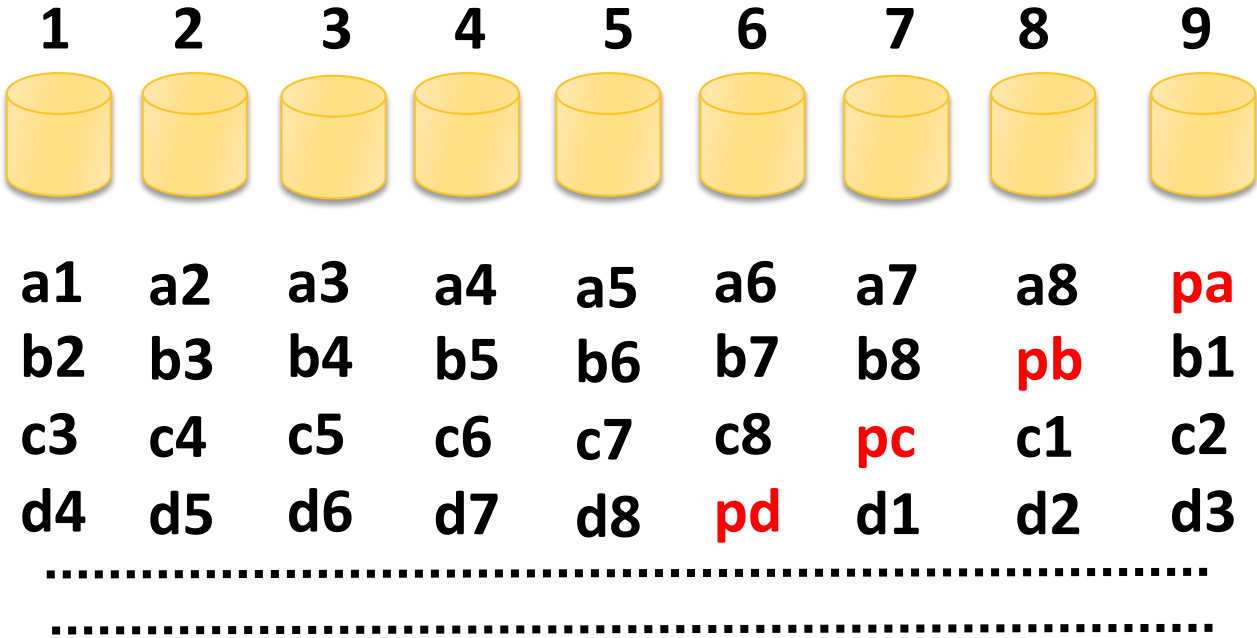
---



Il **RAID 5** ha **1** solo disco in più ma parità e dati sono **distribuiti ciclicamente** su tutti: un blocco di dati viene memorizzato sui dischi da **1 a 8** e la parità sul **9**, un secondo blocco dati viene memorizzato sui dischi da **9 a 7** e la parità sul **8** ....

Il **RAID 6** ha **2** dischi in più ma l'informazione è codificata con il **codice Reed Solomon** e riesce a correggere **due** guasti

# RAID 5



disposizione **Left Symmetric**

# Alcuni concetti sulla probabilità

---

- ▶ Sia  $P(A)$  la probabilità che un evento  $A$  accada in un dato periodo di tempo
- ▶  $P(A)$  è tra 0 e 1 e  $1 - P(A)$  è la probabilità che l'evento  $A$  non accada nel periodo
- ▶ Due eventi  $A$  e  $B$  sono indipendenti se l'occorrenza di uno non influenza l'occorrenza dell'altro, allora:

$$P(A \text{ and } B) = P(A) \times P(B) \text{ (entrambi accadono)}$$

$$P(A \text{ or } B) = P(A) + (1 - P(A)) \times P(B) \text{ (uno solo accade)}$$

$$= P(A) + P(B) - P(A) \times P(B) \approx P(A) + P(B)$$

# Alcuni concetti sulla probabilità

---

- ▶ Poiché le P sono troppo piccole si usa al loro posto il MT(A) cioè il tempo medio di occorrenza dell'evento (*mean time to event*) :  $MT(A) = 1/P(A)$
- ▶ Attenzione, se i tre componenti lavorano insieme MT (almeno uno) è il MT globale! Quindi per alzare il MT globale bisogna alzare la qualità di tutti.

$$MT(G) = 1/(P(A)+P(B)+P(C)) =$$

$$1/(1/MT(A)+1/MT(B)+1/MT(C))$$

se gli MT di N componenti sono uguali allora:

$$MT(G) = MT(A)/N$$

# Analisi della affidabilità

---

- ▶ **Ipotesi:**
  - ▶ i guasti nei vari dischi sono **indipendenti**
  - ▶ la possibilità di guasto è **invariabile** nel tempo
- ▶ **le grandezze di interesse sono :**
  - ▶ **MTTF:** mean time to **failure**
  - ▶ **MTTR:** mean time to **repair**
  - ▶ **MTTDL:** mean time to **data loss**

# Analisi della affidabilità

---

- ▶ **RAID 0** (poco affidabili):

$$MTTF_{RAID0} = MTTF_{DISCO} / N = MTDDL$$

- ▶ **Es.:**

$$MTTF_{DISCO} = 30000 \text{ h} > \mathbf{3 \text{ anni}}$$

con **100 dischi**:

$$MTTF_{RAID0} = 30000 / 100 = 300 \text{ h} \approx \mathbf{2 \text{ sett.}}$$

con **8 dischi** :  $3759 \text{ h} \approx \mathbf{22 \text{ sett.}}$

# Analisi della affidabilità

---

- ▶ **RAID 1** (molto affidabili):

$$MTTF_{RAID1} = MTTF_{DISCO} / (2 \times N)$$

- ▶ **Es.:**

con 16 dischi :  $\approx$  **11 sett.**

**Però!**

**MTTDL** (mean time to data loss) è **elevatissimo**: si dovrebbero guastare contemporaneamente un disco e la sua copia



# Analisi della affidabilità

---

- ▶ **RAID 5** (molto affidabili):

$$MTTDL_{RAID5} =$$

$$(MTTF_{DISCO})^2 / (N \times (N-1) \times MTTR)$$

- ▶ **Es.:**

con 9 dischi e MTTR = 24 h :  $\approx$  **60 anni**

molto inferiore al **RAID 6** ma molto superiore ai dischi **SLED** (single large expensive disk)