

# Acquisizione, memorizzazione e visualizzazione delle informazioni

Davide Quaglia

# Sommario

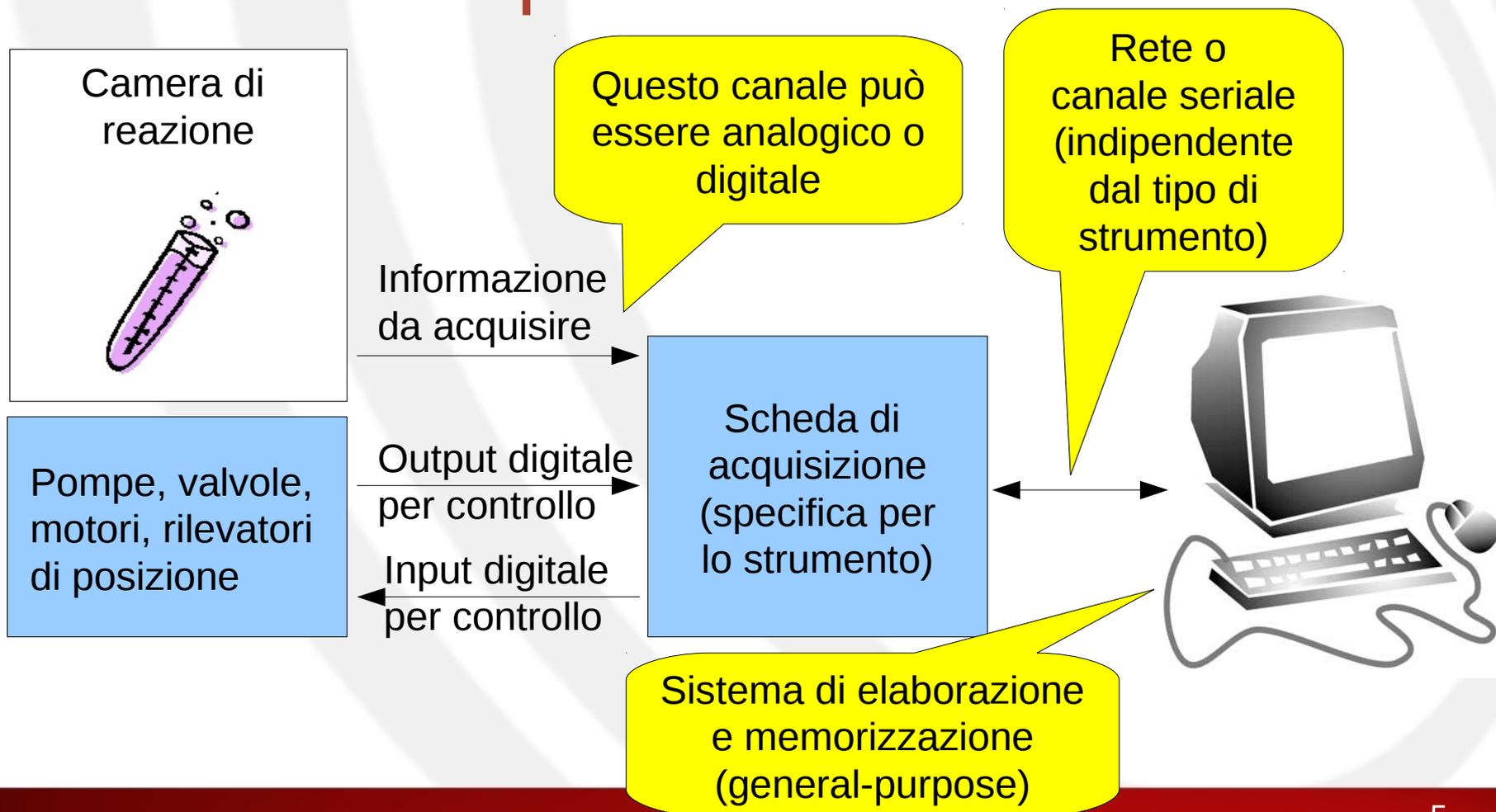
- Acquisizione da sensori
- Cenni sulla compressione
- Memorizzazione su supporti di massa
- Alcuni formati di memorizzazione

# Acquisizione da sensori

# Obiettivo

- Rilevazione/misurazione di un fenomeno bio/chimico/fisico
- Molti strumenti di laboratorio sono oggi automatizzati
  - Occorre trasformare la rilevazione/misurazione in una **sequenza di numeri binari** per essere elaborati e memorizzati da un calcolatore
- Lab-on-chip

# Struttura di una interfaccia computer-strumento



# Natura dell'informazione da acquisire

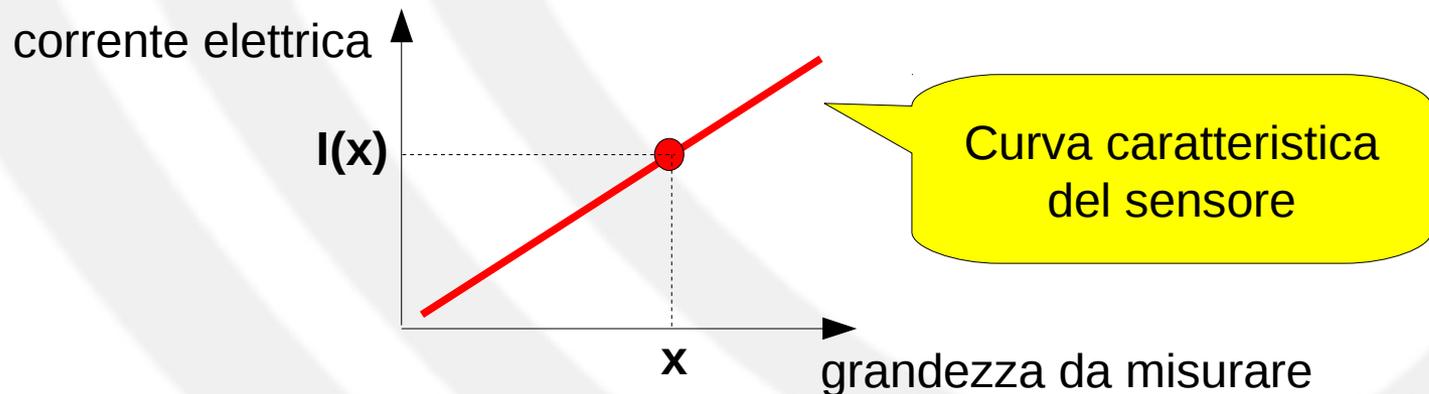
- Caso 1. La rilevazione/misura del fenomeno bio/chimico/fisico consistere nel **conteggio di eventi**
  - l'informazione acquisita è già costituita da numeri
    - Es. il contatore Geiger per la radioattività conta degli impulsi generati da certe particelle nell'unità di tempo
- Caso 2. La rilevazione/misura del fenomeno consiste nella misura di una **grandezza elettrica (tensione o corrente) che varia con continuità** ed è proporzionale all'intensità del fenomeno fisico considerato

# Misura attraverso segnale elettrico

- il sistema di acquisizione misura una **grandezza elettrica** (corrente o tensione)
- occorre una **relazione matematica nota** tra il fenomeno fisico e la sua manifestazione elettrica (curva caratteristica)
- si ottiene un **valore continuo a precisione infinita** che occorre trasformare in un numero attraverso campionamento e quantizzazione (vedi dopo)

# Curva caratteristica

- **Funzione invertibile** che lega la grandezza bio/chimico/fisica da misurare con la grandezza elettrica che viene misurata
- Misura == inversione della funzione



# Fenomeni fisici con manifestazione elettrica

- Intensità luminosa
  - Luce emessa dalla fenomeno stesso
  - Luce trasmessa in trasparenza da una sostanza in funzione del suo grado di torbidità
- Caratteristiche elettriche di atomi o molecole da rilevare/misurare

# Misura di intensità luminosa

- Certi atomi, molecole, microrganismi, se eccitati da energia sotto forma di luce o calore possono emettere luce di una lunghezza d'onda ben precisa (luminescenza, fluorescenza, fosforescenza)
  - Spettroscopia
  - Spettrografia

## Misura di intensità luminosa (2)

- Una soluzione può avere un certo **grado di trasparenza** in funzione della presenza e concentrazione di sostanze o organismi
- Una soluzione può avere un certo **colore** che significa trasparenza selettiva su certe frequenze di luce
- Un reticolo cristallino o una struttura molecolare **regolare** può creare **diffrazione di una luce incidente** (raggi ultravioletti, raggi X)
  - **Diffrattografia**

## Misura di intensità luminosa (3)

- La misura di intensità luminosa può rivelare tali sostanze o determinarne la concentrazione
- I misuratori di intensità luminosa
  - convertono la luce in un segnale elettrico
  - contano eventi luminosi (uscita numerica diretta)
- Spesso le caratteristiche costruttive di tali misuratori li rendono sensibili ad una specifica frequenza dello spettro luminoso
  - Tale selettività si può ottenere anche antepponendo un filtro ottico davanti al misuratore

# Caratteristiche elettriche di atomi o molecole da rilevare/misurare

- Conducibilità elettrica o resistenza elettrica
- Capacità elettrica
- Differenza di potenziale elettrico
- Caratteristiche magnetiche che inducono un campo elettrico

# Conducibilità elettrica

- Legge di Ohm:  $V = R * I$
- R dipendente da:
  - Presenza di ioni → concentrazione ionica
  - Variazione della resistenza in funzione delle caratteristiche fisiche da misurare
    - Termoresistenza: solido metallico la cui resistenza elettrica è proporzionale alla temperatura (Es. PT100 = sonda di platino)

# Misura di concentrazione ionica

- Uno ione è una molecola dotata di carica elettrica
- Il movimento di uno ione e il suo urto contro un metallo genera
  - una corrente elettrica (misurazione di intensità)
  - un impulso elettrico (conteggio)
  - una emissione di luce (misurazione di intensità luminosa)
- Intensità di corrente, numero di impulsi di corrente oppure intensità luminosa sono proporzionali alla concentrazione della sostanza da analizzare
- I dispositivi di misura possono avere un meccanismo di moltiplicazione del flusso di ioni per amplificare l'effetto e facilitarne la rilevazione

# Misura di capacità

- Una soluzione chimica da analizzare posta tra due conduttori crea un condensatore (= una piccola batteria)
  - La capacità può essere valutata misurando il tempo che una corrente elettrica nota impiega a riempire il condensatore
  - Se la capacità del condensatore dipende dalla concentrazione della sostanza cercata allora la misura della capacità permette di risalire alla concentrazione della sostanza

# Differenza di potenziale elettrico

- E' una misura di tensione (si misura in Volt)
- **Termocoppia**: saldatura di due metalli che assume una differenza di potenziale (tensione) proporzionale alla temperatura
- **Sensore piezoelettrico** per misure di pressione e forza: la pressione meccanica su un cristallo crea tra certe sue facce un potenziale elettrico proporzionale alla pressione o forza e misurabile da uno strumento

# Caratteristiche magnetiche

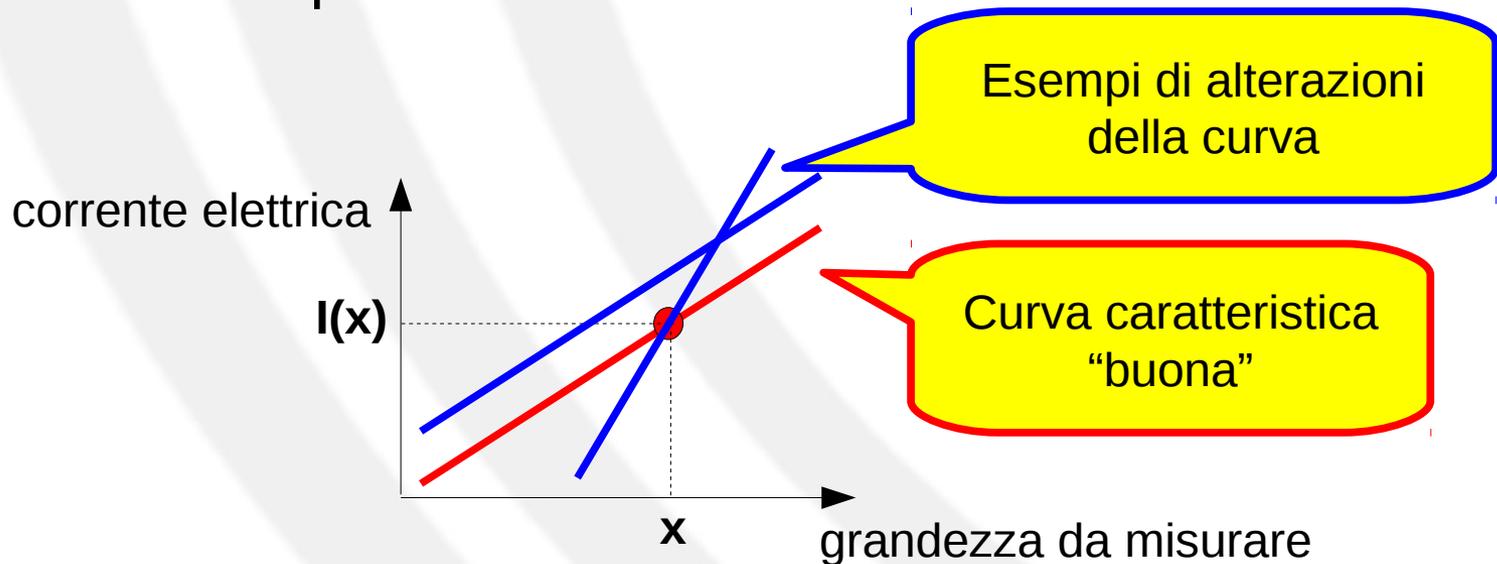
- Certi atomi hanno particolari caratteristiche magnetiche (come fossero tante piccole calamite) e si **caricano di energia** se immersi contemporaneamente
  - in un campo magnetico costante esterno
  - in un'onda radio a frequenza particolare (caratteristica di ogni atomo)
- La rilevazione dell'energia assorbita ad una certa frequenza è proporzionale alla concentrazione dell'atomo corrispondente
- Risonanza magnetica nucleare (NMR)

# Misura indiretta di un fenomeno

- Il fenomeno bio/chimico/fisico da rilevare/misurare si deve sempre manifestare attraverso i meccanismi elencati
- In caso contrario occorre una **misura indiretta**
- Ad es. una molecola senza proprietà luminose/elettriche può essere rilevata da una molecola avente tali proprietà che si lega ad essa (**marcatore**)
  - Micro array, sequenziatore di DNA

# Calibrazione

- La curva caratteristica dei sensori può subire variazioni dovute a fattori ambientali o usura
- Spesso la curva subisce solo una translazione o una variazione di pendenza o di scala

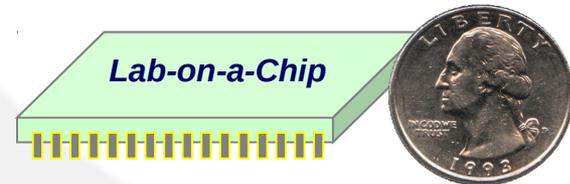


## Calibrazione (2)

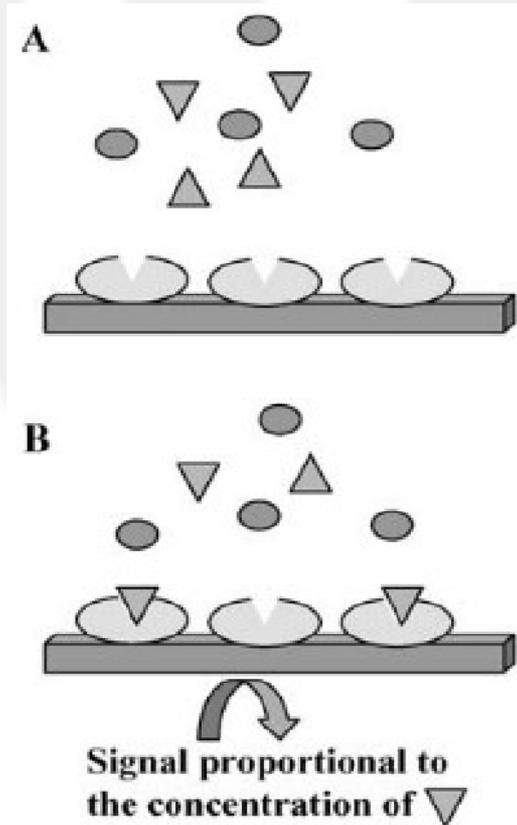
- Prima di iniziare una acquisizione occorre fare una calibrazione
  - si verifica la presenza di una variazione misurando vari campioni di valore noto
  - si compensano le variazioni della curva nel sistema di misurazione (ad es. aggiunta di un costante, moltiplicazione per un fattore correttivo)
  - si lancia un segnale di allarme in caso di compensazione impossibile

# Biosensori e lab-on-a-chip

- Miniaturizzazione del processo di analisi biochimica
- Creare dispositivi economici e veloci, per analisi “vicino al paziente” oppure per monitorare elementi tossici nell'ambiente, nell'acqua e nei cibi
- Consumo di piccole quantità di reagenti e analita; funzionamento automatico, elevata sensibilità



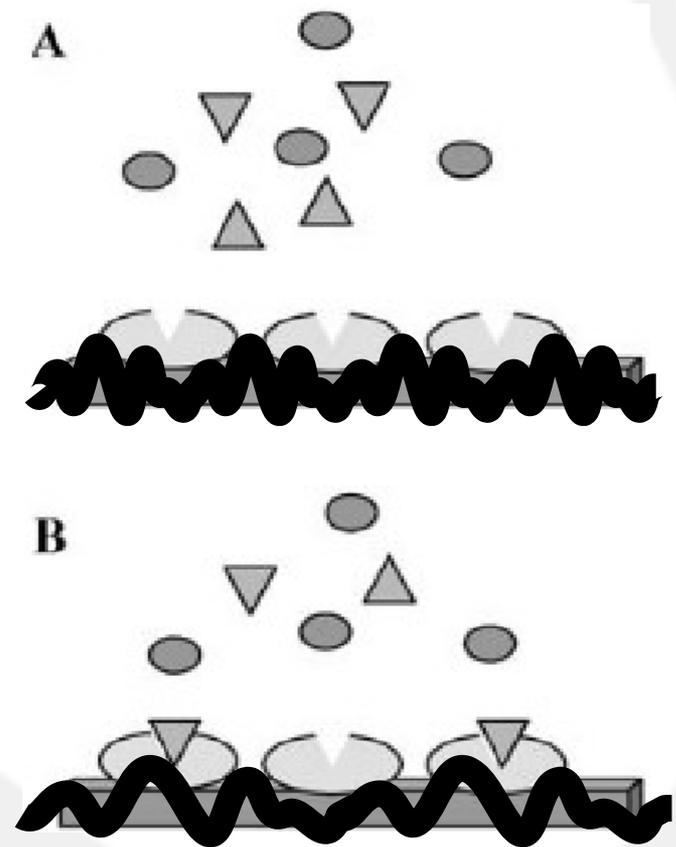
# Biosensori amperometrici



- Prevedono la rilevazione di una variazione di corrente elettrica generata da reazioni di ossidoriduzione catalizzate da enzimi di ossidoreduttasi
- Opportuni nanomateriali forniscono supporto stabile agli enzimi e facilitano la conduzione elettrica

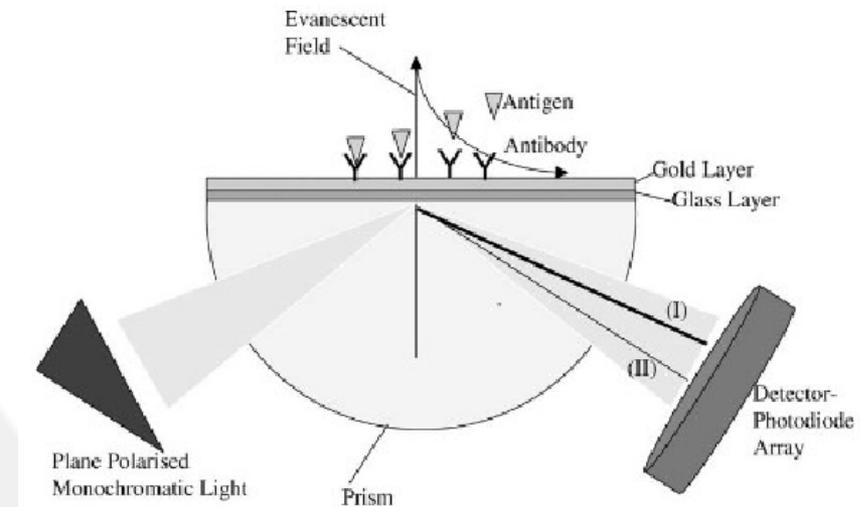
# Biosensori acustici

- Il substrato ricevente contiene molecole in grado di combinarsi con la molecola da rilevare
- Durante l'analisi il sistema substrato+analita ha una massa maggiore e proporzionale alla concentrazione dell'analita
- Il sistema viene fatto oscillare
- La variazione di massa provoca una variazione della frequenza di risonanza

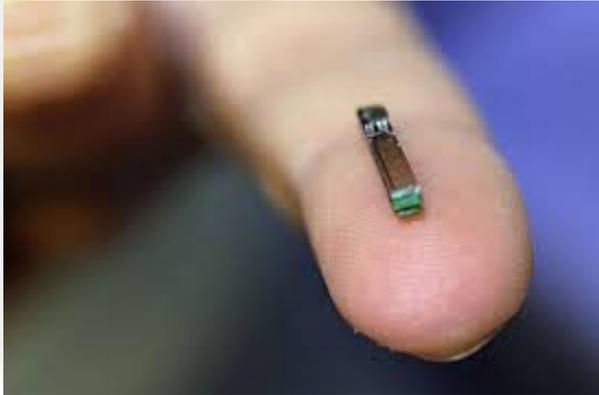


# Biosensori ottici

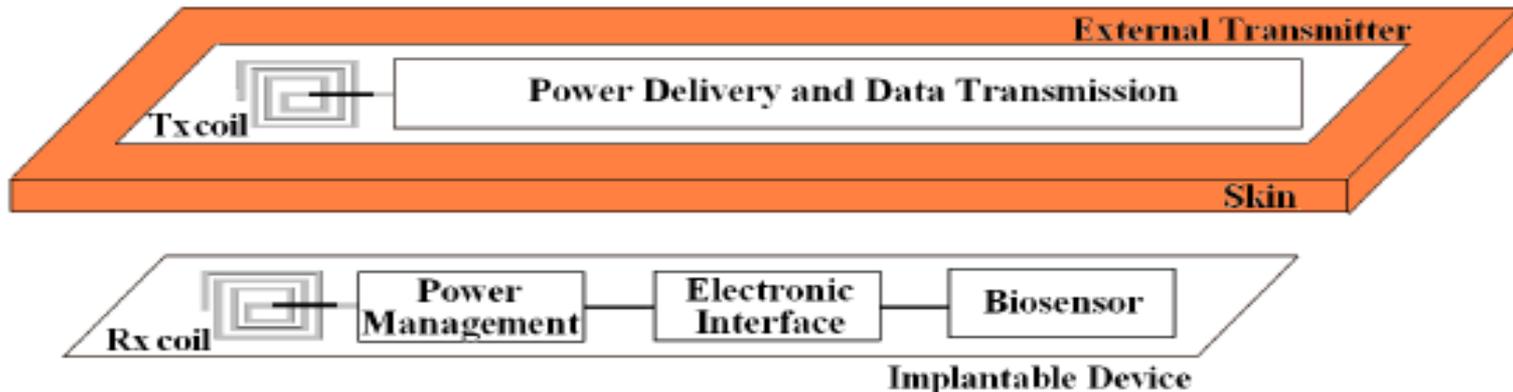
- La combinazione dell'analita col substrato crea una variazione dell'angolo di riflessione della luce polarizzata incidente
- Tale variazione viene rilevata con una matrice di rilevatori di luce (la stessa di una fotocamera digitale)



# Biosensori impiantabili



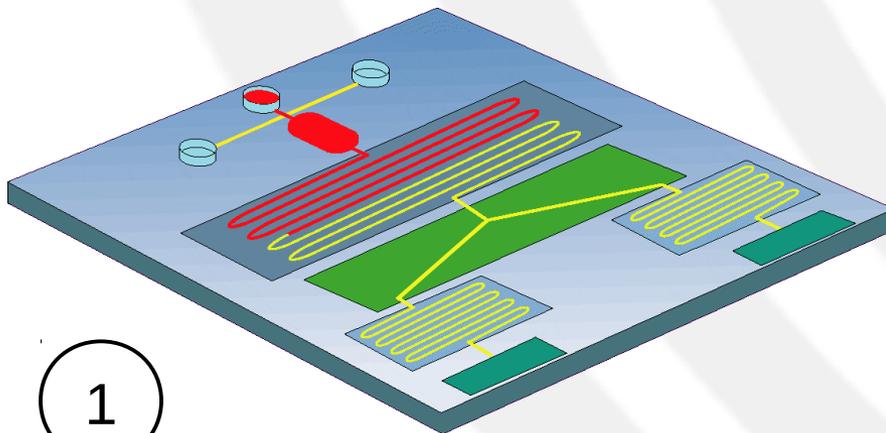
- Due parti:
  - Parte esterna alimentata a batteria per alimentare la parte interna e ritrasmettere info verso la rete locale
  - Parte interna alimentata mediante campo elettromagnetico, dotata di sensore/i, CPU
- Biosensore di tipo amperometrico
- La parte interna deve essere contenuta in un involucro bio-compatibile



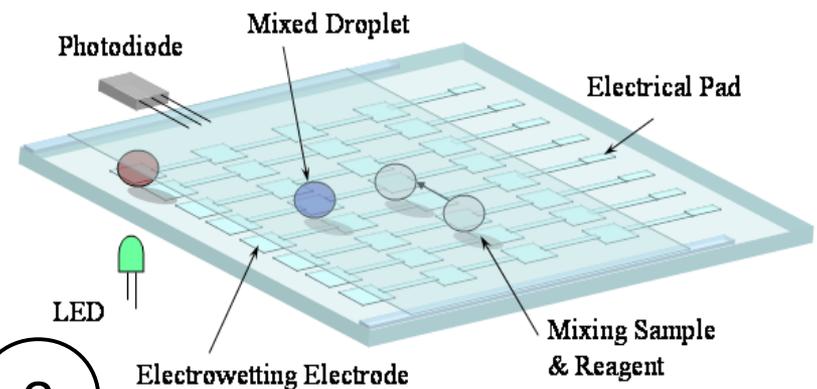
# Microfluidica

Microtubi che movimentano nanolitri di analita, solvente e reagente

- 1) **Continuous-flow lab-on-chip**: micropompe e microcanali
- 2) **Digital microfluidic lab-on-chip**: creazione di gocce (droplets) e loro spostamento/miscelazione tramite impulsi elettrici



1



2

# Segnale analogico e segnale digitale

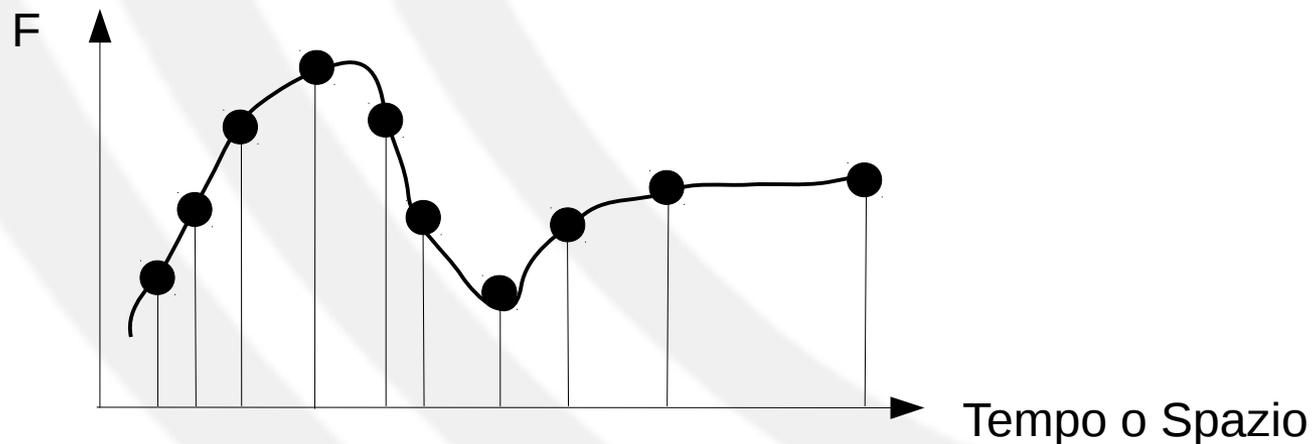
- Un **segnale analogico** consiste di **valori continui** esprimibili solo mediante **numeri reali**
- Un **segnale digitale** consiste in una **sequenza discreta** di numeri interi con **precisione finita**
  - i calcolatori sanno elaborare solo segnali digitali e, in particolare, rappresentati in base 2

# Conversione analogico/digitale

- La conversione analogico/digitale consiste in due azioni **indipendenti**:
  - campionamento
  - quantizzazione

# Campionamento di un segnale

- Per catturare la variazione di un segnale nel suo dominio (es. tempo o spazio) occorre ripetere la **misura** in punti ad intervalli regolari di tale dominio (es. istanti di tempo o punti dello spazio).
- Intuitivamente più il segnale varia velocemente (ad es. nel tempo o nello spazio) e maggiormente vicini devono essere tali punti cioè maggiore deve essere la **frequenza di campionamento**.



# Quanti campioni ?



16 x 21 pixel

# Quanti campioni ?



64 x 85 pixel

# Quanti campioni ?



128 x 171 pixel

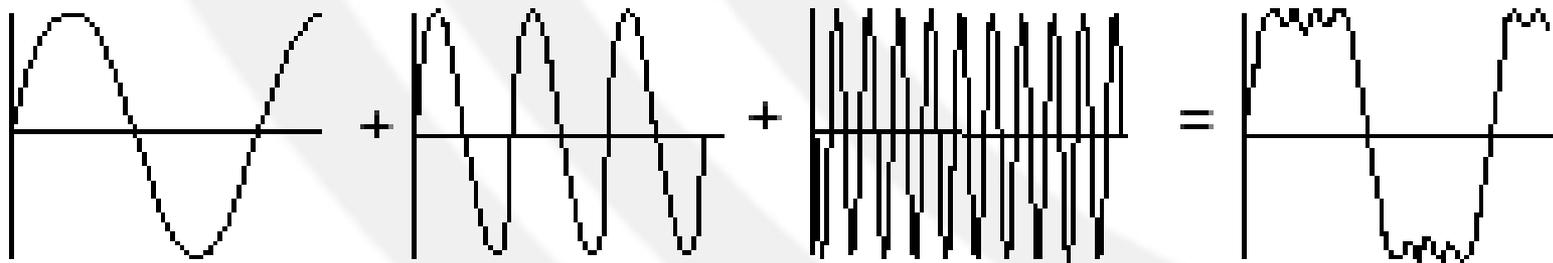
# Quanti campioni ?



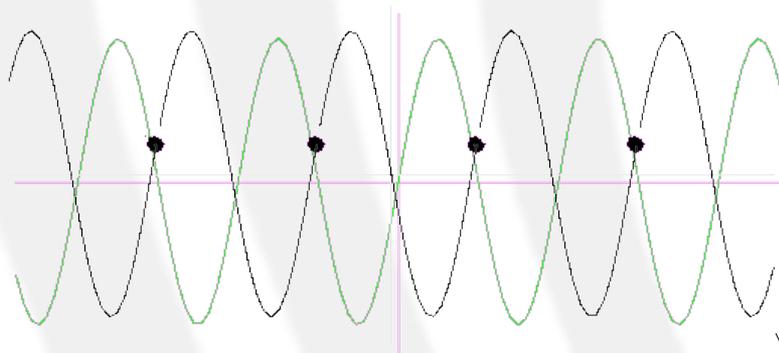
512 x 683 pixel

# Teorema del campionamento

- Un segnale può sempre essere rappresentato come **somma di sinusoidi** (rappresentazione in frequenza).
- Per ricreare fedelmente un segnale occorre ricreare fedelmente **tutte** le sinusoidi che lo compongono.



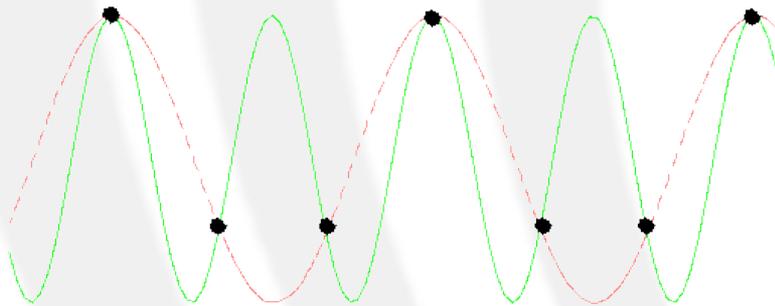
# Teorema del campionamento (2)



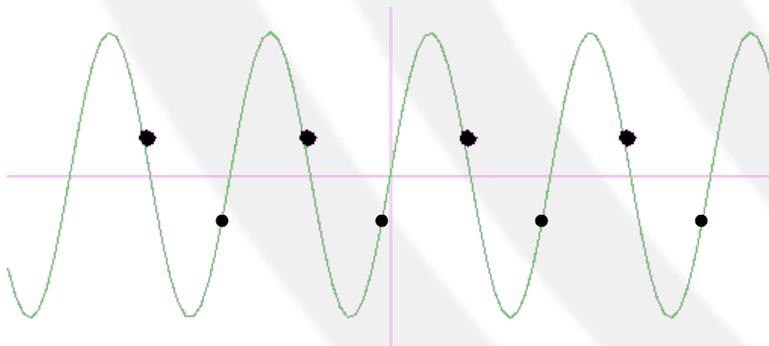
1 campione per periodo

aliasing

# Teorema del campionamento (3)



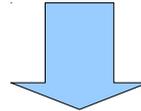
1.5 campioni per periodo



**2 campioni per periodo**

# Teorema del campionamento (4)

- Per riprodurre fedelmente una sinusoide occorrono **2 campioni per periodo.**



- **La frequenza di campionamento deve essere doppia della frequenza della sinusoide.**
- Se una certa frequenza di campionamento mi permette di rappresentare fedelmente la sinusoide con frequenza  $f$  allora la stessa frequenza di campionamento rappresenta fedelmente tutte le sinusoidi con frequenza minore di  $f$ .

# Teorema del campionamento (5)

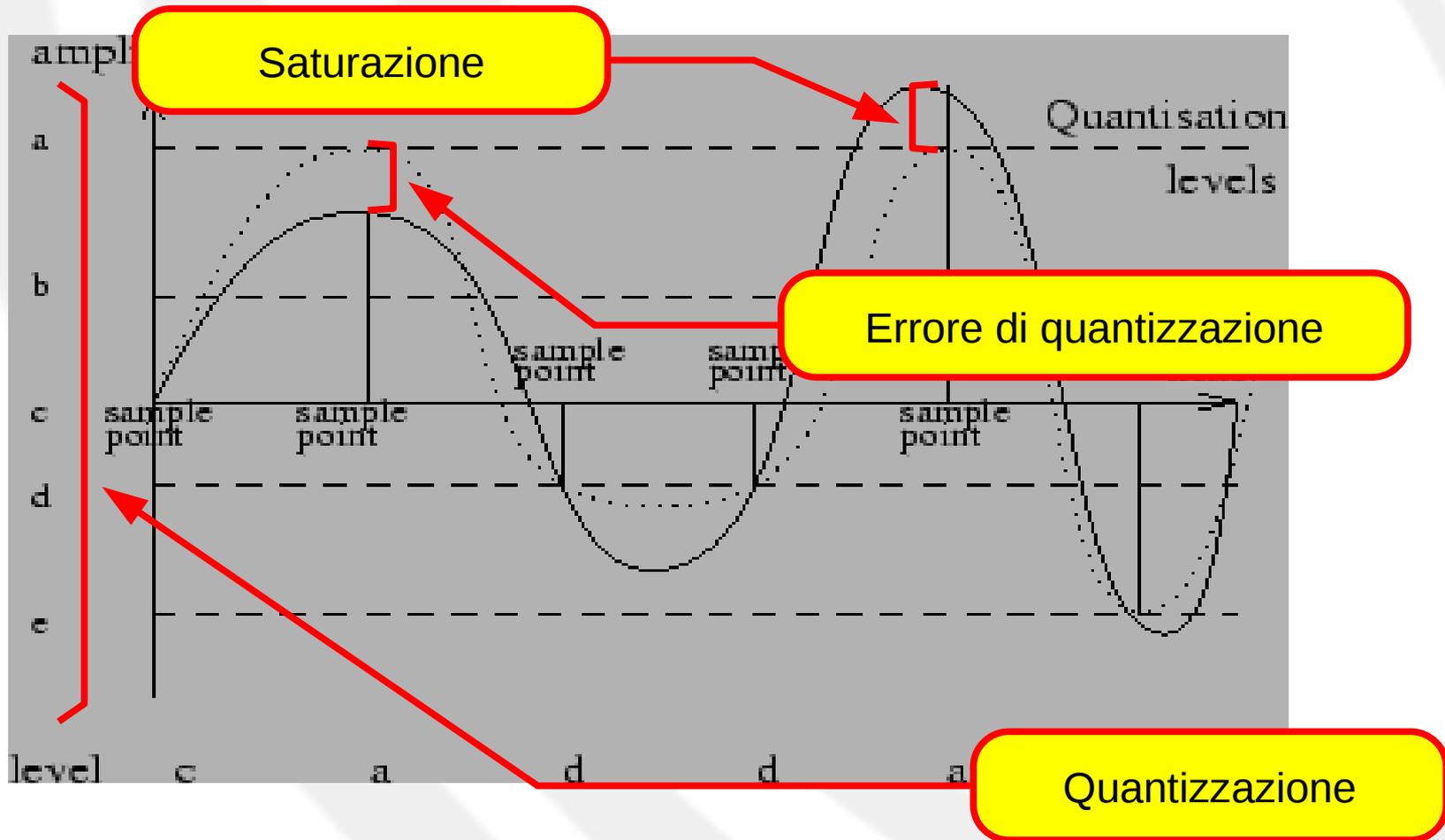
- Tutti i segnali continui naturali possono considerarsi composti da un numero finito di sinusoidi
- Se la sinusoide a frequenza più alta ha frequenza  $f$  allora si dice che il segnale ha **larghezza di banda  $f$**
- Quindi la frequenza di campionamento di tale segnale deve essere **doppia della larghezza di banda del segnale (Teorema di Nyquist)**.

# Quantizzazione di una misura

Assegnazione di una misura a precisione infinita ad un numero avente precisione finita.

- 1) Determinazione di un range di variazione (detta **dinamica del segnale**).
- 2) Individuazione di un **insieme finito** di livelli all'interno di tale range.
- 3) Assegnamento di **un numero intero** a ciascun livello.
- 4) Misurazione del segnale analogico e rappresentazione della misura con il numero corrispondente al livello più vicino.

# Quantizzazione: modalità e problemi



# Effetto della quantizzazione sulle immagini



256 livelli (8 bit)



8 livelli (3 bit)

# Effetto della quantizzazione sulle immagini



256 livelli (8 bit)



8 livelli (3 bit)

L'effetto osservato non ha nulla a che vedere con la frequenza di campionamento che è uguale nei 2 casi !

# Effetto della quantizzazione

- La quantizzazione implica sempre un errore di arrotondamento detto di **errore di quantizzazione**.
- Maggiore è il numero di livelli (e quindi di bit) su cui si rappresenta la misura e minore è l'errore di quantizzazione.
- Tale errore è irreversibile.
  - Quando un procedimento di acquisizione/memorizzazione dati prevede uno o più fasi di quantizzazione allora non è possibile ritornare perfettamente alla rappresentazione originale.  

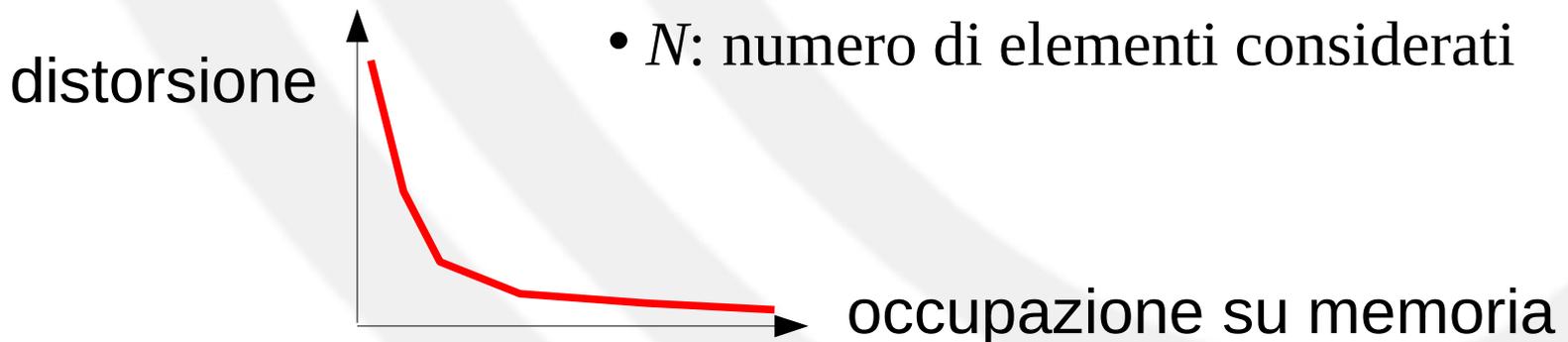
  - **Codifica dell'informazione con perdita (lossy)**

# Distorsione di quantizzazione

- Una possibile misura dell'errore di quantizzazione è detta **distorsione**
- Una possibile misura della distorsione su una sequenza  $\{v_i\}$  è data da:

$$MSE = \frac{\sum_1^N (v_i - v'_i)^2}{N}$$

- $v_i$ : simbolo prima della quantizzazione
- $v'_i$ : simbolo dopo la decodifica
- $N$ : numero di elementi considerati



# Cenni sulla compressione delle informazioni

# Premessa

- Quanto si dirà in questi lucidi si applica a sorgenti di informazioni che producono **serie di dati discreti a precisione finita**:
  - dati digitali su file (es., eseguibili, database)
  - sorgenti naturali dopo campionamento e quantizzazione
- Le tecniche che verranno presentate **non** provocano una perdita irreversibile di informazione.

# Definizioni

- Una **sorgente** di informazione emette una **serie ordinata di dati**
- Tale serie di dati si può rappresentare come una **sequenza di simboli** dove ciascun simbolo appartiene ad certo **alfabeto** dipendente dal tipo di applicazione
  - Es: campioni di temperatura quantizzati su 16 bit con segno
    - l'alfabeto è costituito dai numeri interi nell'intervallo  $[-32768, 32767]$
  - Es: mappa di fluorescenza con pixel su 256 livelli di grigio
    - l'alfabeto è costituito dai numeri interi nell'intervallo  $[0,255]$

# Cosa vuol dire compressione ?

- Premesso che ogni simbolo deve essere codificato con dei bit per essere elaborato o trasmesso
- Definizione teorica di compressione

***Cambio reversibile di alfabeto tale che la rappresentazione binaria della sequenza di simboli, tradotta nei simboli del nuovo alfabeto, richiede meno bit per essere memorizzata.***

# Eliminazione della correlazione statistica

# Correlazione statistica

- Sia data una sequenza ordinata di simboli  $s_0, s_1, \dots, s_n$  di un certo alfabeto emessa da una sorgente di informazione.
- C'è correlazione quando la probabilità che  $s_i$  assuma un certo valore **dipende** dal valore assunto dagli elementi precedenti della sequenza.
- Se c'è correlazione statistica, si dice che l'informazione contenuta nella sequenza è **ridondante** e si può cambiare alfabeto in modo da usare meno bit.

# Esempio di correlazione nei testi

- 33% di lettere eliminate
- quante parole si riesce a ricostruire ?

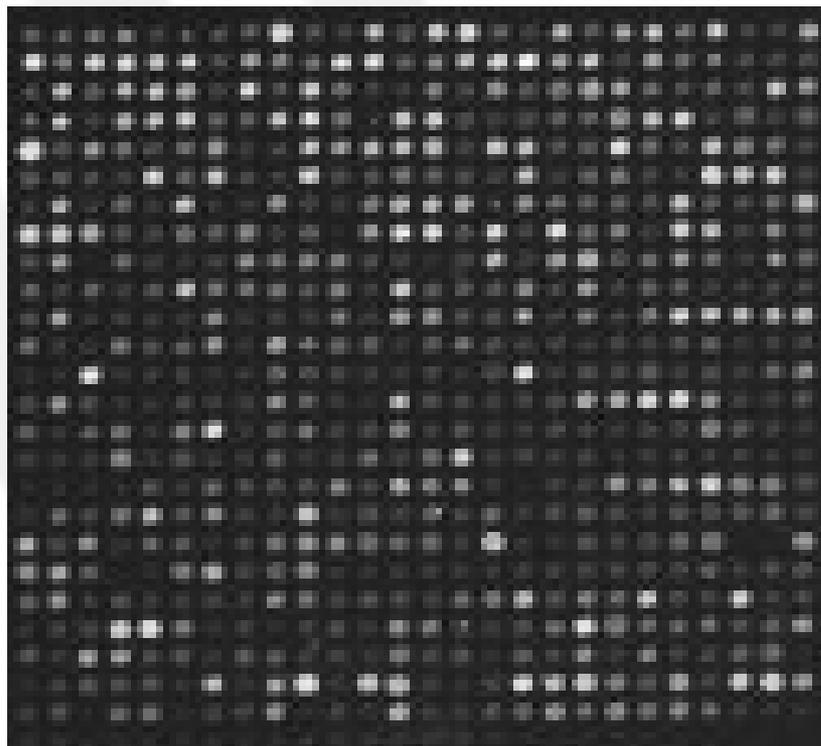
Il mio cognome è Gallo e il  
 mio animale domestico  
 preferito è il gallo. Puoi  
 chiamarmi al numero  
 telefonico 0580781

# Esempio di correlazione nei testi

Il mio cognome è Gatto e il  
mio animale domestico  
preferito è il gatto. Puoi  
chiamarmi al numero di  
telefono 0858027821

# Esempio di correlazione nelle immagini

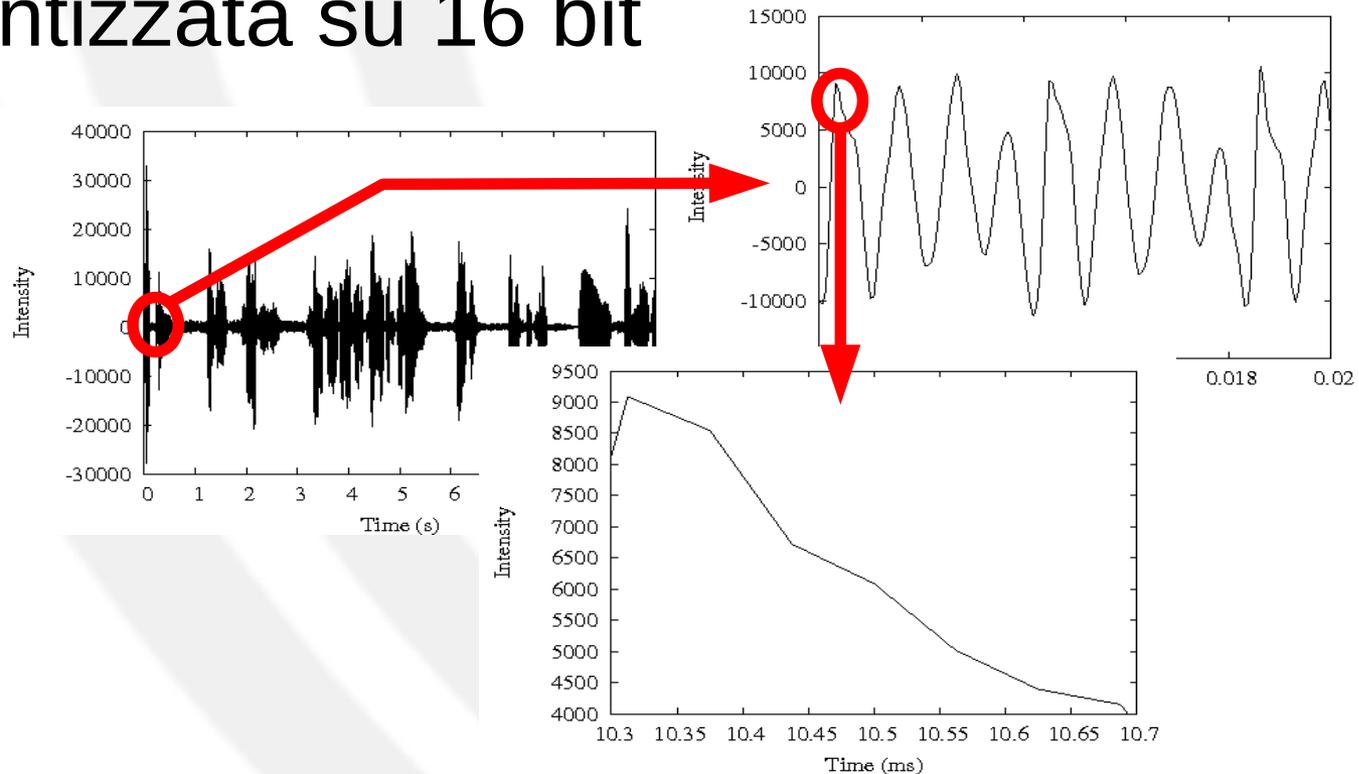
Immagine b/n da scansione di micro array



← Spazi tra un pozzetto e l'altro

# Esempio di correlazione in una sequenza temporale

- Andamento nel tempo della temperatura quantizzata su 16 bit



# Eliminazione della correlazione

- **Si cambia alfabeto di rappresentazione**

- Codifica run-length

- Es in una immagine b/n a due livelli

1,1,1,1,0,0,0,0,0,0,1,1,1,1,1 --> 4,6,5 (assumendo di iniziare sempre col nero)

- Codifica differenziale

- Es. della temperatura (vedi lucido precedente)

$s_0 s_1 s_2 \dots s_n \rightarrow (s_1 - s_0), (s_2 - s_1), \dots, (s_n - s_{n-1})$

- Trasformata

- Es. della temperatura (vedi lucido precedente)

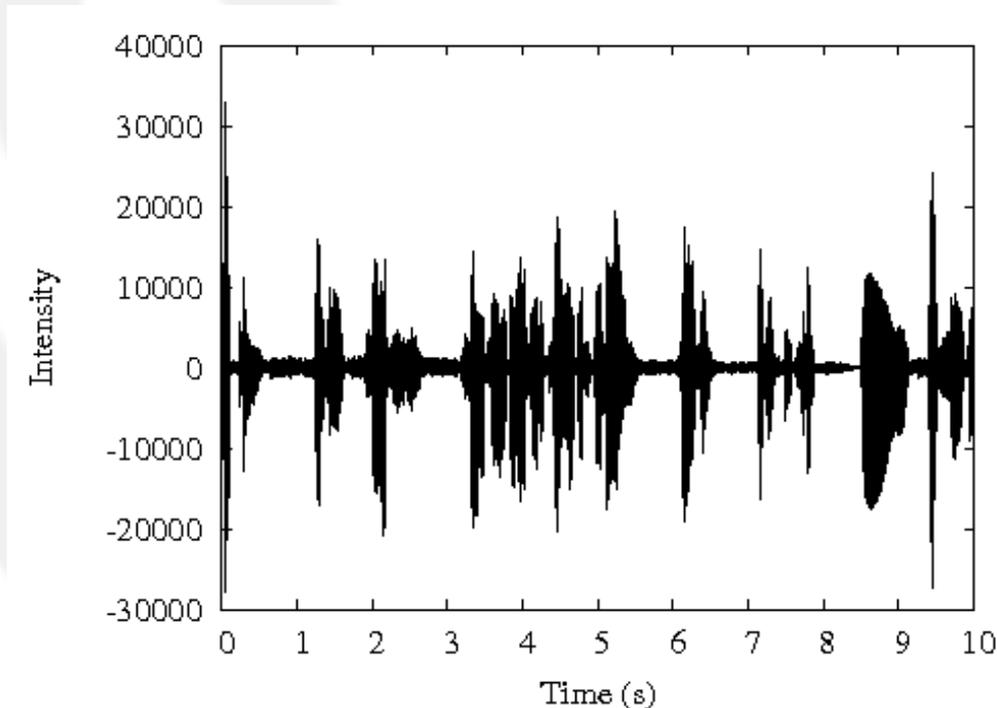
Sviluppo in serie di Fourier: ampiezza delle componenti sinusoidali

$s_0 s_1 s_2 \dots s_n \rightarrow a_0, a_1, \dots, a_n$

# Codifica entropica

# Frequenza dei simboli: esempio

- Sequenza temporale su 16 bit: alfabeto = interi su  $[-32768, 32767]$



# Quanti bit/simbolo ?

- Dato un certo alfabeto  $A$  contenente  $N$  simboli, quanti bit/simbolo?
- Al 1° anno è stato detto  $\lceil \log_2 N \rceil$
- In realtà se in una sequenza  $S$  rappresentassi con meno bit i simboli più frequenti e con più bit quelli meno frequenti la lunghezza in bit della sequenza sarebbe minore (grazie Morse!).
- Allora esiste un numero medio ottimale di bit/simbolo che però dipende dalla statistica della sorgente  $S$  considerata

## Quanti bit/simbolo ? (2)

- Dato un alfabeto  $A$  di simboli, il numero medio ottimale di bit/simbolo per rappresentare la sorgente  $s$  è:

$$H(s) = \sum_{k \in \text{alfabeto } A} p_k \log_2 \frac{1}{p_k}$$

- $p_k$  è la frequenza di apparizione del simbolo  $k$ -esimo in  $s$
- **$H(s)$  è l'entropia della sorgente  $s$  (Shannon)**
- Se usassi tale valore, la lunghezza in bit della sequenza sarebbe minima senza perdere informazione

# Esempio di calcolo dell'entropia

- Una sorgente emette 100 simboli appartenenti ad un alfabeto di 4 simboli  $\{a, b, c, d\}$
- Frequenze di apparizione dei 4 simboli sono:
  - $1/2, 1/4, 1/8, 1/8$  (NB: somma delle frequenze = 1)
- Entropia della sorgente:

$$H(s) = \frac{1}{2} \log_2(2) + \frac{1}{4} \log_2(4) + \frac{1}{8} \log_2(8) + \frac{1}{8} \log_2(8)$$

$$= \frac{1}{2} + \frac{1}{2} + \frac{3}{8} + \frac{3}{8} = 1.75 \text{ bit / simbolo} \leq \lceil \log_2(4) \rceil \text{ bit / simbolo}$$

- Dimensione con approccio 1° anno:  $100 \cdot 2 = 200$  bit
- Dimensione con approccio nuovo: 175 bit

# Codifica entropica

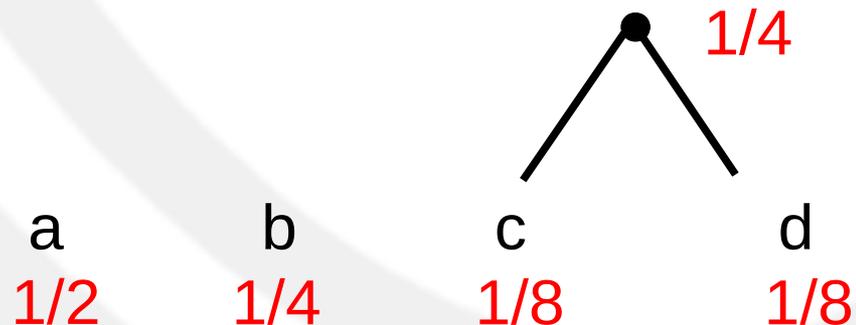
- Per ottenere una lunghezza media dell'alfabeto uguale o vicina al valore dell'entropia occorre rappresentare i simboli su un numero variabile di bit
- Problema 1: come si fa a costruire la rappresentazione binaria ottima ?
- Problema 2: come si fa a sapere quando inizia/finisce un simbolo ?

# Algoritmo di Huffman

- Metodo di assegnamento bit ai simboli che garantisce la corretta decodifica
- Si costruisce un albero binario dove le foglie rappresentano i simboli
- Ad ogni biforcazione la somma delle frequenze dei simboli del sotto-albero di destra deve essere il più possibile vicina alla somma delle frequenze dei simboli del sotto-albero di sinistra.
- Ciascun arco di una biforcazione è etichettato con 0 e 1 rispettivamente

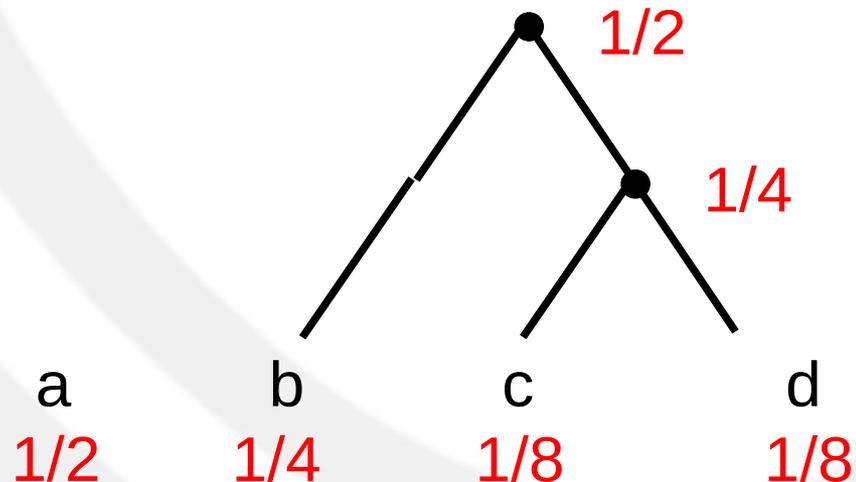
# Creazione codice di Huffman

- Alfabeto di 4 simboli  $a, b, c, d$  con frequenze rispettivamente di  $1/2, 1/4, 1/8, 1/8$
- Primo passo:  $a, b, c, d$   
 $1/2 \quad 1/4 \quad 1/8 \quad 1/8$



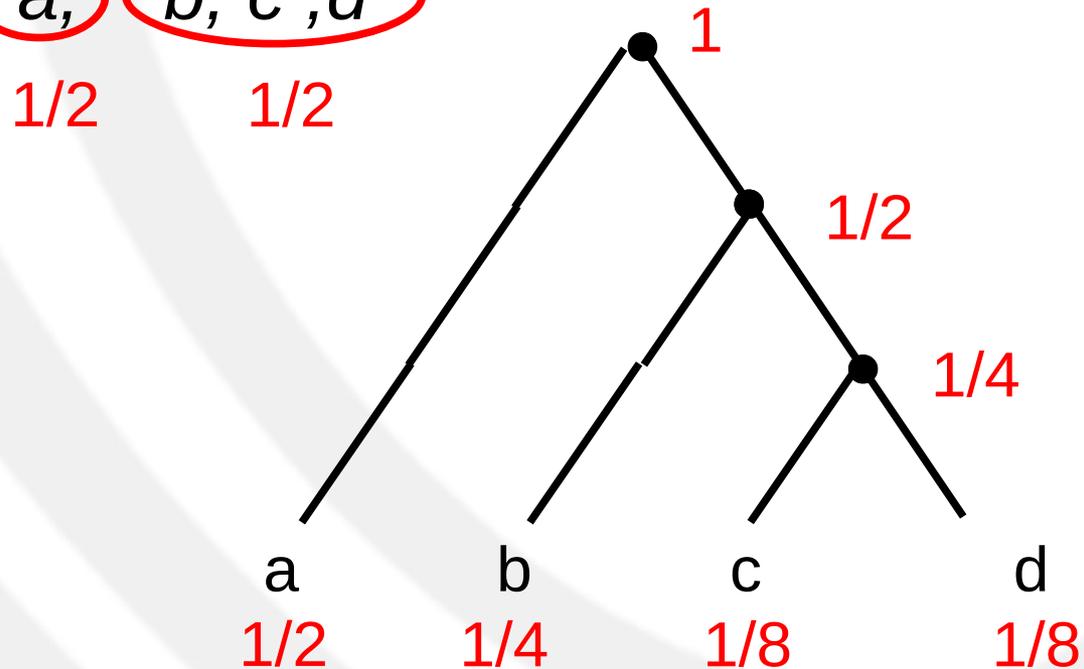
# Creazione codice di Huffman

- Alfabeto di 4 simboli  $a, b, c, d$  con frequenze rispettivamente di  $1/2, 1/4, 1/8, 1/8$
- Secondo passo:  $a, b, c, d$   
 $1/2 \quad 1/4 \quad 1/4$



# Creazione codice di Huffman

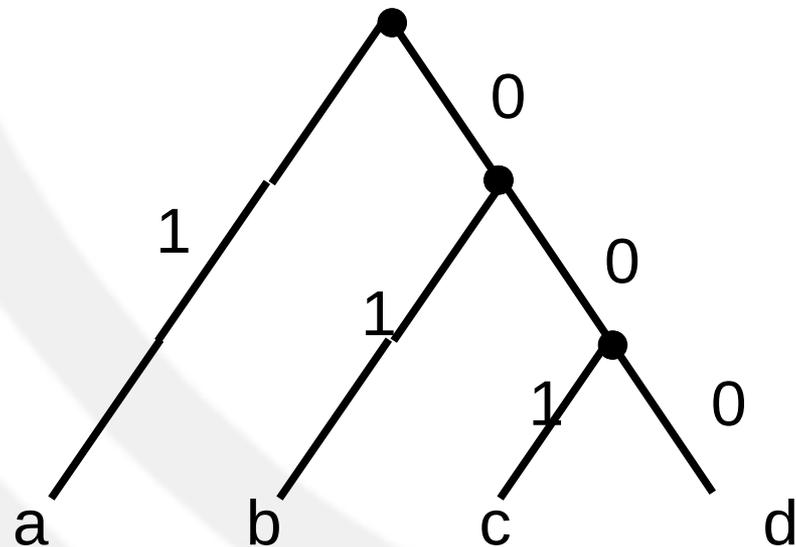
- Alfabeto di 4 simboli  $a, b, c, d$  con frequenze rispettivamente di  $1/2, 1/4, 1/8, 1/8$
- Terzo passo:  $a,$   $b, c, d$



# Creazione codice di Huffman

- Alfabeto di 4 simboli  $a, b, c, d$  con frequenze rispettivamente di  $1/2, 1/4, 1/8, 1/8$
- Creazione del codice:

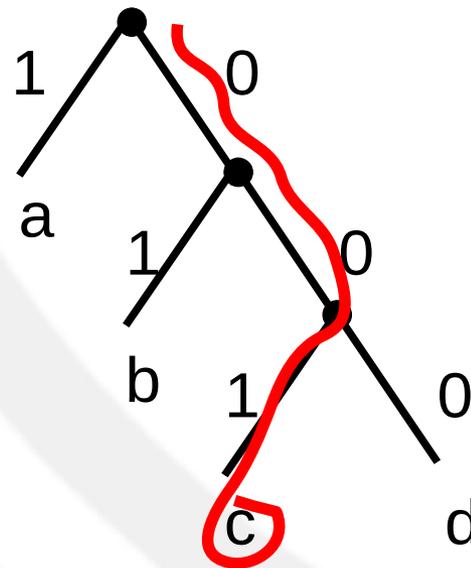
$a \rightarrow 1$   
 $b \rightarrow 01$   
 $c \rightarrow 001$   
 $d \rightarrow 000$



# Decodifica codice di Huffman

- In ricezione per decodificare i simboli si visita l'albero binario in base ai bit in arrivo

Es: 001 ---> c



# Huffman: considerazioni

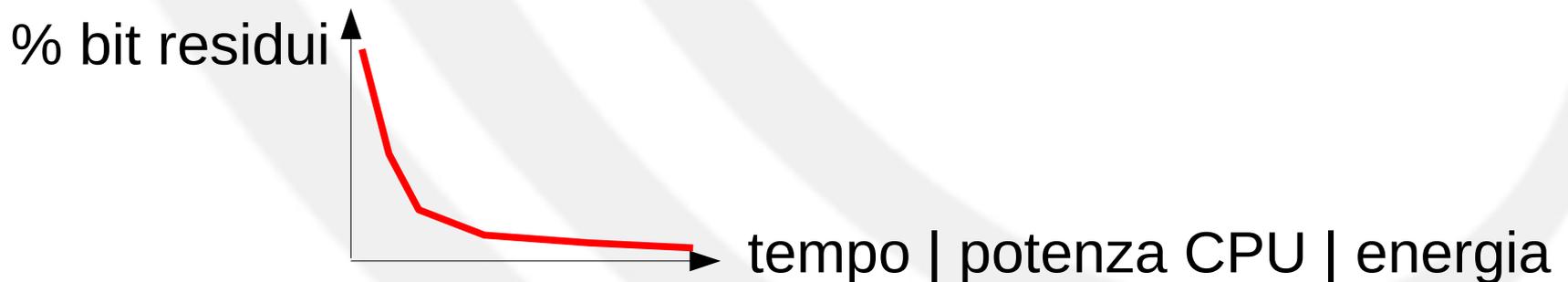
- L'algoritmo di Huffman è ottimo (rispetto all'entropia) se le frequenze sono potenze di 2.
- Per distribuzioni di frequenze più complicate esistono algoritmi più sofisticati
  - Codifica aritmetica
  - Codici di Golomb

# Codifica entropica: altri problemi

- Stima delle frequenze  $p_k$ 
  - Migliore è la stima e maggiore è la compressione
  - Siccome la statistica dipende dalla sorgente bisognerebbe progettare una codifica ad-hoc per ogni sorgente
    - Poco pratico
  - Occorre fare ipotesi vincolanti sul tipo di sorgente oppure considerare un elevato numero di esempi

# Considerazioni sulla complessità

In generale se si vuole comprimere di più si devono usare **algoritmi più complessi** che impiegano più tempo oppure necessitano di una **CPU più potente** e comunque portano sempre ad un **maggiore consumo di energia**.

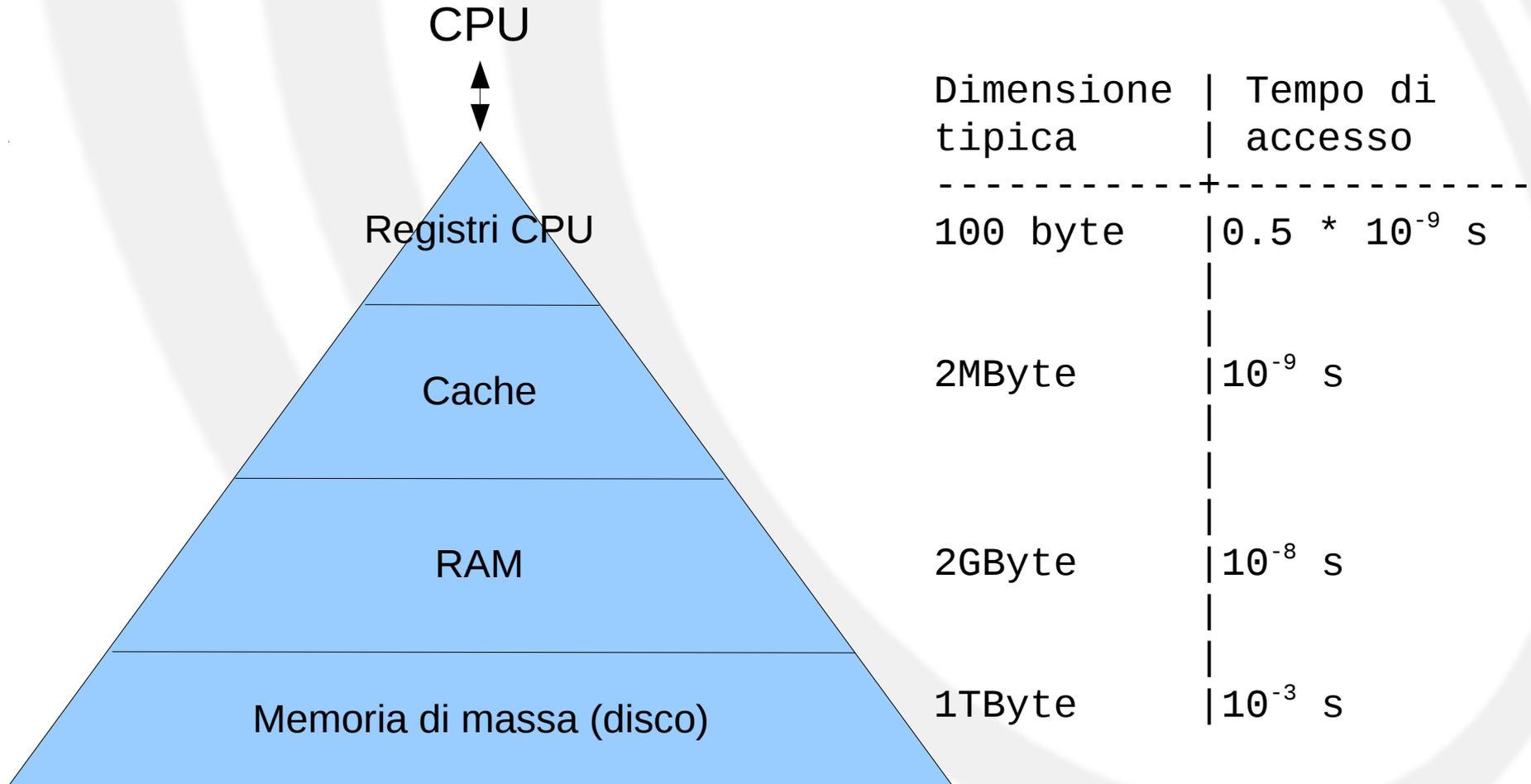


# Memorizzazione su supporti di massa

# Metriche di prestazioni

- Capacità:
  - $1\text{ K} = 2^{10}$   $1\text{ M} = 2^{20}$   $1\text{ G} = 2^{30}$   $1\text{ T} = 2^{40}$
- Tempo di accesso: tempo che intercorre tra la richiesta di un dato da parte della CPU e l'arrivo del dato alla CPU

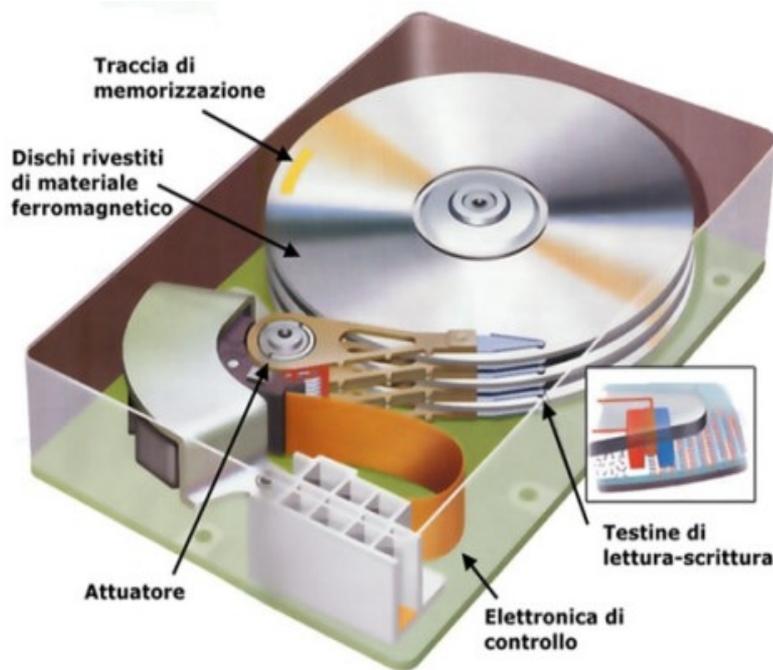
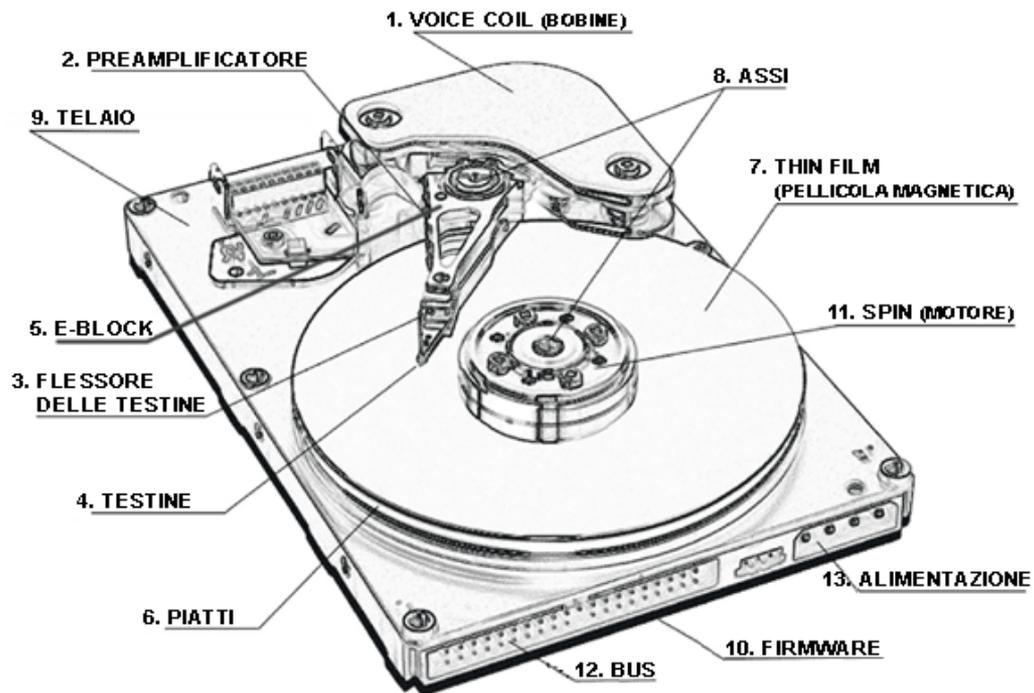
# Gerarchia di memoria



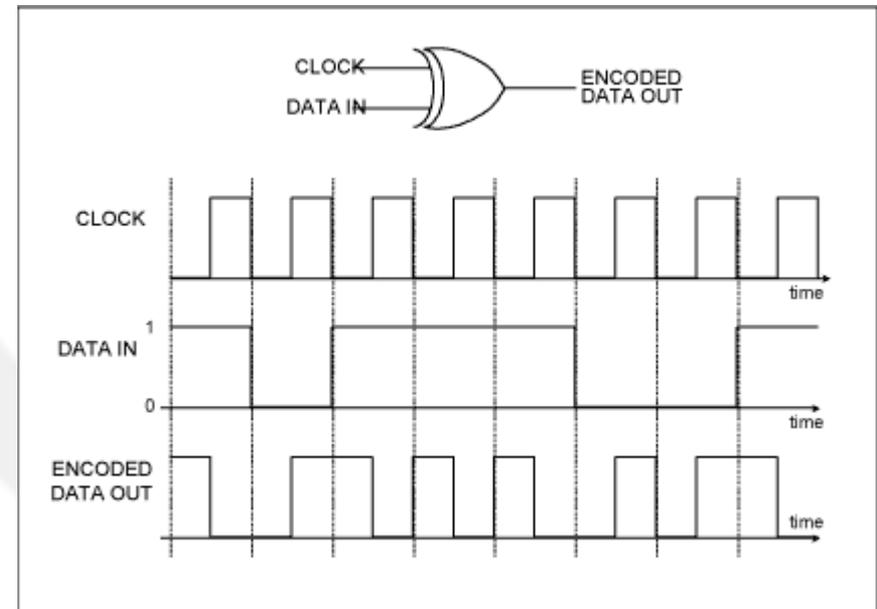
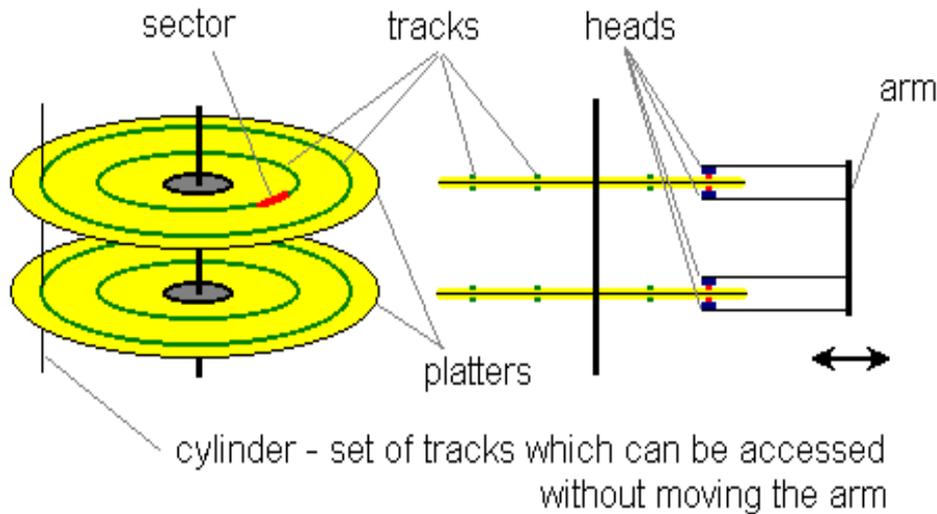
# Memoria di massa

- **Funzione:**
  - ospitare dati e programmi in modo permanente
  - serbatoio per la memoria virtuale
- **Tecnologie**
  - disco magnetico: il più diffuso
  - disco ottico
  - memoria flash (a stato solido): si sta diffondendo

# Disco magnetico



# Principio di funzionamento



# Organizzazione dei dati

- Il **settore** è l'unità minima di lettura/scrittura
  - **identificato univocamente da**
    - numero di faccia
    - numero di traccia
    - numero di settore all'interno della traccia
  - **contiene 512 o 1024 byte**
- Il sistema operativo si occupa di creare la *metafora* del File System (file, cartelle, link) mentre comanda al disco **letture e scritture a quantità intere di settori**

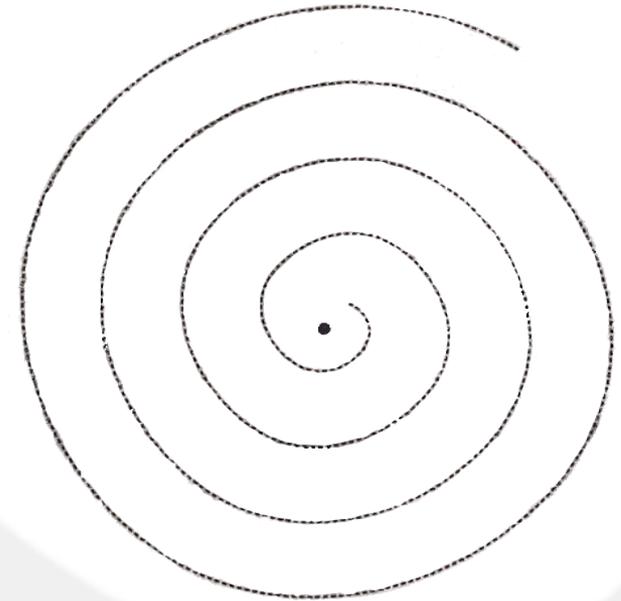
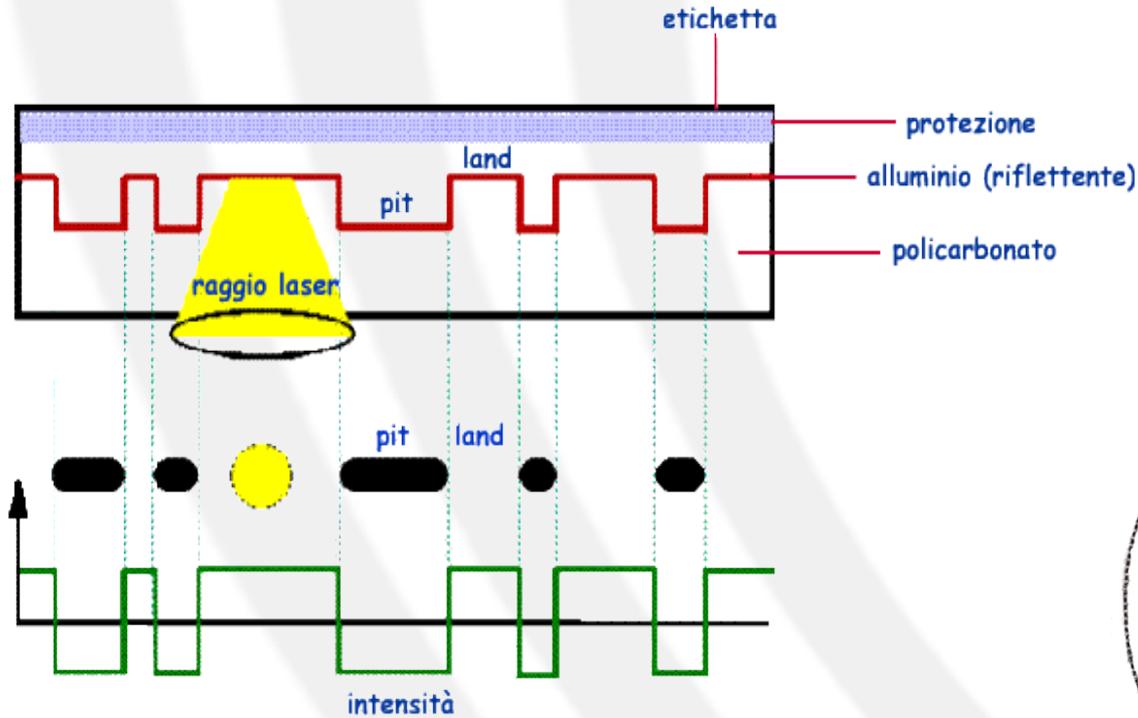
# Componenti del tempo di accesso

- Tempo di seek = tempo per posizionare il braccio meccanico sulla traccia (in media 5-10 ms)
- Latenza rotazionale = tempo di attesa prima che il settore scelto passi sotto le testine di lettura/scrittura
  - In media  $1/2 * 1/V_{\text{rotazione}}$ 
    - A 10000 giri/minuto  $1/2 * 60/10000 = 3 \text{ ms}$
- Tempo di trasferimento del settore
  - Se la velocità di trasferimento è 20-40 MB/s per trasferire un settore di 512 byte ci vogliono  $512/V_{\text{transfer}} = 12-24 \mu\text{s}$
- Il tempo di seek è dominante

# Evoluzione dei dischi magnetici



# Dischi ottici



# Dischi ottici

- CD
  - 700 MB
- DVD
  - Singolo lato, singolo strato (4.7 GB)
  - Singolo lato, doppio strato (8.5 GB)
  - Doppio lato, singolo strato (9.4 GB)
  - Doppio lato, doppio strato (17 GB)
- Blu-ray
  - Minore lunghezza d'onda del laser e quindi maggiore densita'
  - 25 GB per lato

# Memoria a stato solido o flash

- Memoria a semiconduttore come la RAM
- I bit vengono memorizzati sotto forma di carica elettrica in microscopici condensatori
- Non perde l'informazione quando si toglie l'alimentazione (a differenza della RAM)
- Usata per chiavette USB, memorie per fotocamere, dischi a stato solido, apparati di laboratorio (subito pronti all'accensione)

# Dischi a stato solido

- Vantaggi
  - rumorosità assente
  - minore possibilità di rottura
  - minori consumi
  - tempo di accesso ridotto: decimi di ms
  - maggiore resistenza agli urti
  - minore produzione di calore
- Svantaggi
  - un maggiore costo per bit rispetto a quelli magnetici
  - minore durata del disco a causa del limite di riscritture delle memorie flash.

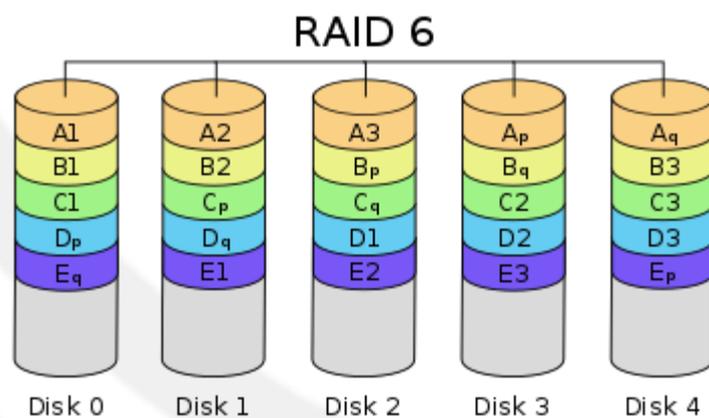
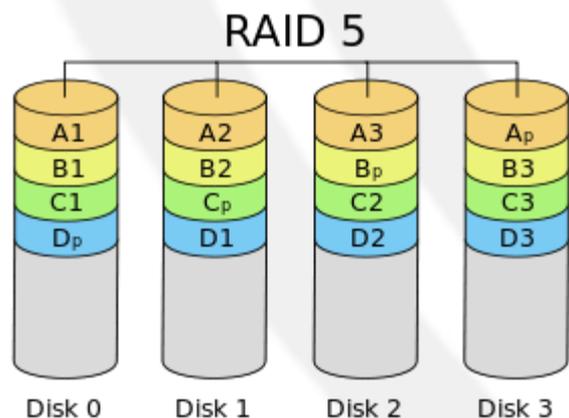
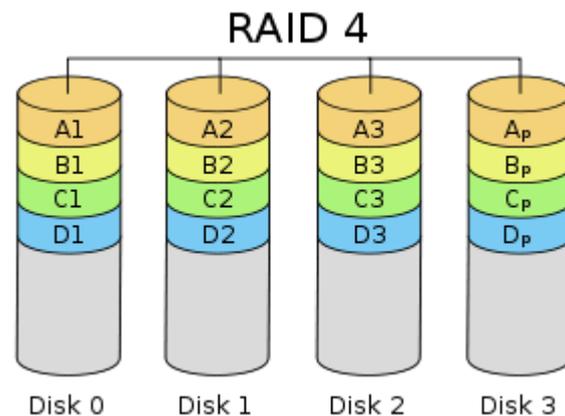
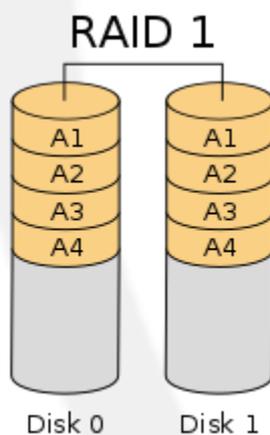
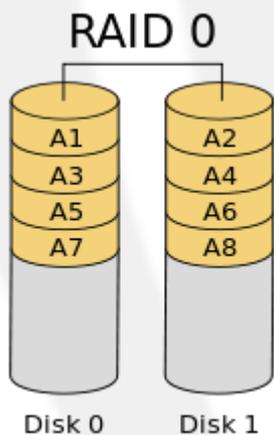
# Redundant Array of Inexpensive/Independent Disks (RAID)

- Motivazione
  - CPU veloce ma dispositivi di I/O lenti
  - Il costo dei dischi è relativamente basso
- Soluzione: utilizzare in maniera coordinata un **insieme di dischi fissi** che però vengono visti dalla CPU come un unico disco
  - Si può aumentare la velocità di I/O parallelizzando le operazioni
  - Si può aumentare l'affidabilità del sistema introducendo ridondanza nella memorizzazione

# RAID: architettura

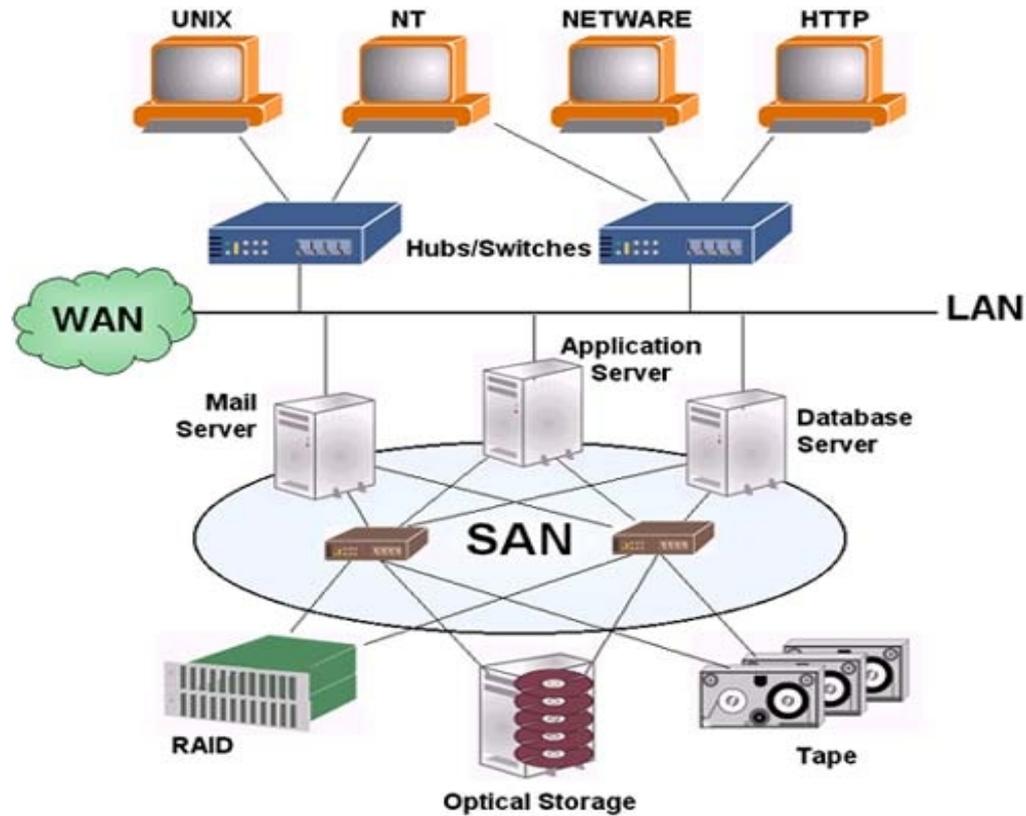
- Insieme dei dischi
- Un controllore RAID viene installato tra la CPU e il sistema di dischi per mascherarne la molteplicità alla CPU
- Il controllore può essere implementato
  - in HW (maggiori prestazioni)
  - In SW come driver del sistema operativo

# RAID: possibili configurazioni



# Storage Area Network (SAN)

## Storage Area Networks



Source: allSAN Report 2001

Copyright © 2000 allSAN.com Inc

**allSAN**.com

# Ruolo del File System

- Il sistema operativo astrae dalle caratteristiche fisiche dei supporti di memorizzazione...
  - Tecnologia di memorizzazione
  - Lettura/scrittura di blocchi di byte
- ... fornendone una visione logica
  - File, collegamenti, cartelle
- File = spazio virtuale di memoria continuo identificato da un nome e atto a contenere un insieme di informazioni sequenziali
  - Tipi di file
    - Dati: testuali, binari
    - Programmi

# Attributi di un file

- Attributi sono memorizzati su disco nella struttura della cartella che lo contiene
  - **Nome del file**
  - **Tipo**
  - **Posizione**
    - Puntatore allo spazio fisico sul dispositivo di memoria
  - **Dimensione**
  - **Protezione**
    - Diritti di lettura/scrittura/esecuzione
  - **Tempo, data e identificazione dell'utente/gruppo possessore del file**

# Operazioni su file

- **Creazione**
  - Cercare spazio su disco, nuovo elemento nella cartella contenente per attributi
- **Scrittura**
  - System call che specifica nome file e dati da scrivere
  - Occorre creare il file oppure cercare nuovo spazio fisico per un file esistente
- **Lettura**
  - System call che specifica nome file e dove mettere dati letti in memoria
  - Occorre scandire il contenuto del file in maniera appropriata
- **Riposizionamento all'interno di file**
  - Aggiornamento puntatore posizione corrente
- **Cancellazione**
  - Libera spazio associato al file e l'elemento corrispondente nella cartella
- **Troncamento**
  - Mantiene inalterati gli attributi ma cancella contenuto del file

# Operazioni su file (2)

- **Apertura**

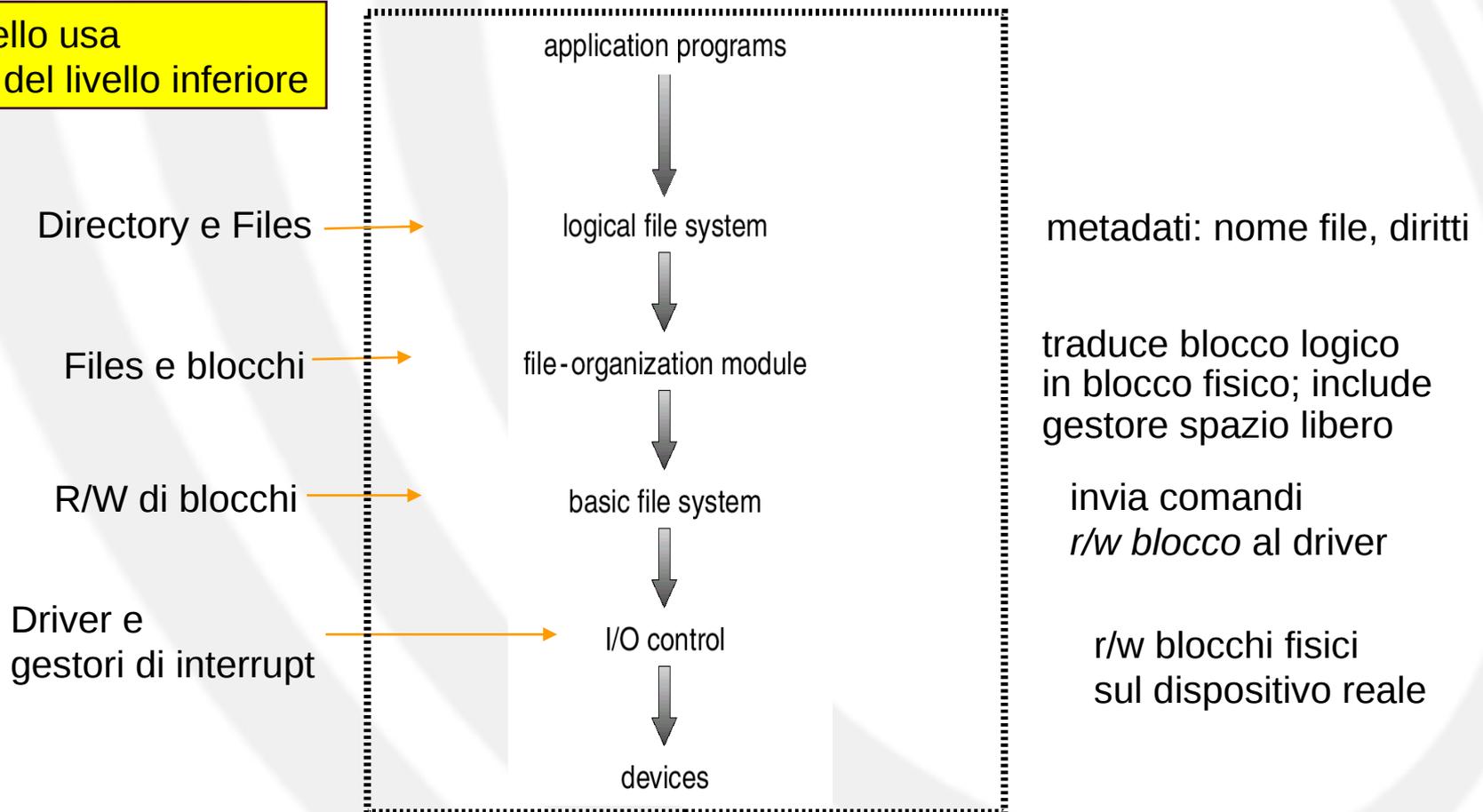
- Ricerca del file nella struttura della directory su disco, copia del file in memoria RAM e inserimento di un riferimento nella tabella dei file aperti
  - In sistemi multiutente ci sono 2 tabelle per gestire i file aperti
    - Una tabella per ogni processo contiene riferimenti per file aperti relativi al processo (es.: puntatore alla locazione di lettura/scrittura)
    - Una tabella per tutti i file aperti da tutti i processi contiene i dati indipendenti dal processo (es.: posizione sul disco, dimensione file, data accessi, n° di processi che hanno aperto il file)

- **Chiusura**

- Copia del file dalla memoria RAM nel disco

# File System a livelli

Ogni livello usa funzioni del livello inferiore



# Implementazione del file system

- Per gestire un file system si usano diverse strutture dati
  - Parte sulla memoria di massa in questione
  - Parte in memoria RAM
- Caratteristiche fortemente dipendenti dal sistema operativo e dal tipo di file system
  - Esistono caratteristiche comuni a tutte le tipologie

# Strutture su memoria di massa

- **Blocco di boot**
  - Informazioni necessarie per l'avviamento del S.O.
- **Blocco di controllo delle partizioni**
  - Dettagli riguardanti la partizione
    - Numero e dimensione dei blocchi, lista blocchi liberi, lista descrittori liberi, ...
- **Strutture di directory**
  - Descrivono l'organizzazione dei file
- **Descrittori di file (inode)**
  - Vari dettagli sui file e puntatori ai blocchi dati

# Strutture in RAM

- **Tabella delle partizioni**
  - Informazioni sulle partizioni *montate*
- **Strutture di directory**
  - Copia in memoria delle directory a cui si è fatto accesso di recente
- **Tabella globale dei file aperti**
  - Copie dei descrittori di file
- **Tabella dei file aperti per ogni processo**
  - Puntatore alla tabella precedente ed informazioni di accesso

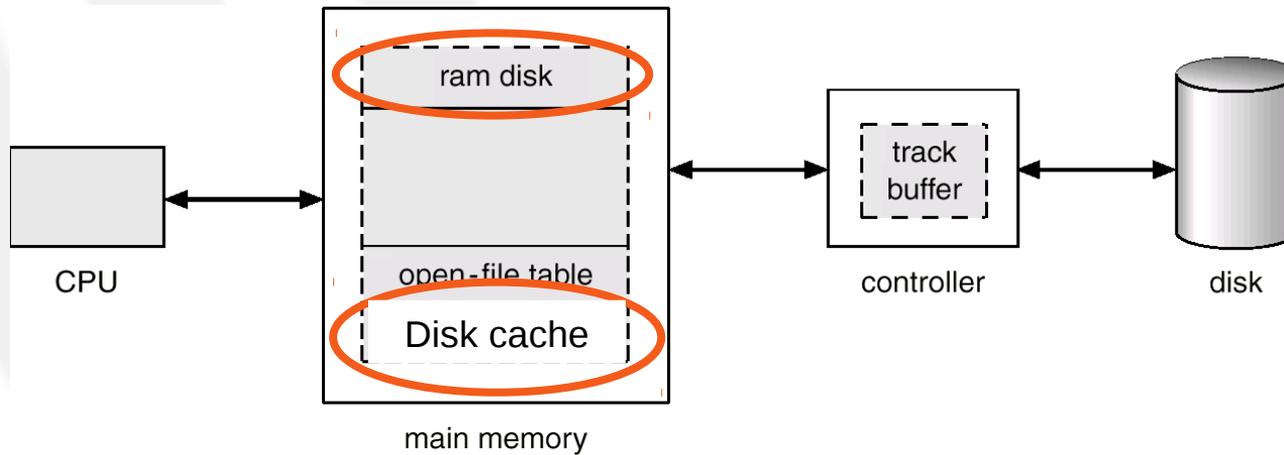
# Efficienza e prestazioni

- La memoria di massa è il collo di bottiglia per le prestazioni del sistema
- Efficienza dipende da
  - Tempo di seek, vel. rotazione, vel. trasferimento dati verso la CPU
  - Algoritmo di allocazione dello spazio su disco
    - Distribuzione dei blocchi di un file sulla superficie del disco
      - continui zig-zag del braccio meccanico fanno perdere tempo

## Efficienza e Prestazioni (2)

- Il controller del disco possiede una piccola cache che è in grado di contenere un'intera traccia ma...
- ... non basta per garantire prestazioni elevate, quindi
  - Dischi virtuali (RAM disk)
  - Cache del disco (detta anche buffer cache)

# Efficienza e Prestazioni (3)



# RAM disk

- Non ha nulla a che vedere con i concetti di **memoria virtuale** e **virtualizzazione**
- **Definizione: parte della memoria RAM viene fatta vedere alle applicazioni come se fosse un disco**
- Il driver di un RAM disk accetta tutte le operazioni standard dei dischi eseguendole però in memoria RAM
  - Veloce
  - Ha senso solo per file temporanei: se spengo perdo tutto!
- L'applicazione decide se scrivere sul RAM disk invece che sul disco vero e proprio nel momento in cui sceglie una ben precisa directory (es. /tmp/)
  - Comportamento diverso dal caso della cache (vedi dopo) che è totalmente trasparente all'applicazione

# Cache del disco

- Porzione di memoria RAM che memorizza blocchi usati di frequente
  - Simile alla cache tra memoria e CPU
  - Gestita dal S.O. e trasparente alle applicazioni
- Sfrutta principio della località
  - Spaziale
    - uso di dati “vicini” a quelli appena usati
  - Temporale
    - uso a breve tempo degli stessi dati
- Le operazioni di read/write su file del processo utente non richiedono uno spostamento di byte sul disco fisico
  - Diminuzione dei tempi di attesa

# Cache del disco: problematiche

- Politica di rimpiazzamento: cosa fare in caso di necessità di eliminare un settore dalla RAM?
  - Least Recently Used, Least Frequently Used, random, ecc...
- Politica di scrittura: se l'operazione è una scrittura, come aggiornare il contenuto su disco?
  - Write-back: scrivo solo quando devo rimuovere il settore dalla cache
    - Problemi di affidabilità in caso di crash
  - Write-through: scrivo sempre
    - Meno efficiente, la cache fa risparmiare tempo solo in lettura

# Problemi di file system in caso di guasto

- Crash del sistema: possibile inconsistenza tra disco e cache
- Guasto del disco: possibile inconsistenza su disco tra i dati nella directory e i dati relativi ai file
  - utilizzo di RAID
  - utilizzo di file system con log (journaled file system)
  - *backup* periodico del disco su memoria di massa ad alta capacità (nastri)
    - recupero di file persi tramite *restore* dei dati dai backup
    - si perde tutto ciò che è stato fatto dopo l'ultimo backup
  - utilizzo di tool di riparazione
    - scandisk (Windows), fsck (Unix)
    - Va fatto appena avviene il guasto altrimenti potrebbe essere troppo tardi !

# Journalled file system

- Registra ogni cambiamento del file system come una *transazione*
  - Tutte le transazioni sono scritte su un *file di log (giornale o diario)*
    - Le transazioni sul log sono scritte in un secondo tempo su disco a causa della cache
  - Quando il contenuto del disco è modificato concordemente alla transazione, questa viene cancellata dal log
  - Se il sistema va in crash, le transizioni non avvenute sono quelle presenti sul log
- Vantaggio
  - Tolleranza ai guasti di sistema
  - Ottimizzazione dell'accesso al disco
- Esempio: file system ext3 di Linux

# Alcuni formati di memorizzazione

# Formati per immagini

- Modi standard di memorizzare e trasmettere immagini digitali
  - Entertainment
  - Immagini mediche
  - Risultati di esperimenti
- Tali formati possono essere
  - basati su pixel o vettoriali
  - lossy o lossless

# Formati basati su pixel (raster o bitmap)

- Si basano su una descrizione dell'immagine come matrice bidimensionale di valori di colori (picture element = pixel)
- Risoluzione (spaziale) = numero di righe x numero di colonne rapportato alle dim reali (es. cm o pollici) sul supporto di visualizzazione o stampa
- Profondità di colore = numero di bit per ciascun pixel per rappresentare il suo colore:
  - Spazi di colore
    - RGB
    - YUV, YCbCr

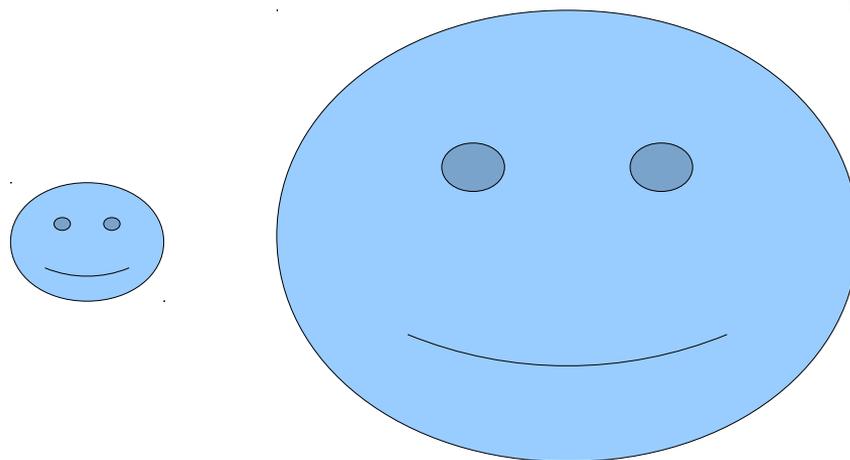
## Formati vettoriali

- L'immagine è descritta come composizione di elementi grafici di base (primitive grafiche)
  - Segmenti, curve
  - Punti
  - Cerchi, ellissi
  - Poligoni
- Per ogni elemento è definita posizione e colore

# Formati raster vs. vettoriali



Raster



Vettoriale

# Formati lossless e lossy

- Spesso la rappresentazione dell'immagine prevede una fase di compressione
- Compressione lossless
  - la fase di compressione non pregiudica la qualità dell'immagine (risoluzione, colore) quando verrà de-compressa
  - utilizza tecniche di rimozione della ridondanza statistica e codifica entropica
  - usata per archiviazione e per applicazioni mediche
- Compressione lossy
  - la fase di compressione pregiudica irreversibilmente la qualità dell'immagine (risoluzione, colore)
  - utilizza anche tecniche di rimozione di dettagli poco percepibili
  - usata per applicazione di entertainment
- A parità di immagine di partenza, le tecniche lossy comprimono di più di quelle lossless

# Formati più comuni

- Raster
  - JPEG
  - TIFF
  - GIF
  - PNG
- Vettoriale
  - SVG
  - Postscript
  - PDF
  - SWF
  - Windows Metafile

# Joint Photographic Experts Group (JPEG)

- Tecnica lossy basata sulla trasformata coseno discreta (DCT) per separare le frequenze spaziali
  - eliminazione della ridondanza statistica
  - eliminazione delle frequenze più alte (→ tecnica lossy)
- Profondità di colore: YUV 8 bit per componente
- Se la qualità è scadente compare il fenomeno della blockingness (blocchettizzazione)
- Utilizzato in fotocamere digitali e per codificare le immagini nei documenti pdf

# Tagged Image File Format (TIFF)

- Permette sia compressione lossless sia lossy
- Profondità di colore: 8 o 16 bit per componente
  - Scale di grigio
  - RGB
  - CMYK
  - CIELab
- Utilizzato
  - dagli scanner (es. fotografie di microarray)
  - frequentemente nel mondo tipografico
  - da alcune fotocamere per evitare il degrado del jpeg

# Graphics Interchange Format (GIF)

- Lossless
- Palette indicizzata a 8 bit (256 colori contemporaneamente nell'immagine)
  - In caso di sfumature si usa il dithering
- Formato adatto a memorizzare immagini sintetiche con larghe zone di colore uniforme come grafici, loghi, ecc...
- Memorizza più immagini in sequenza permettendo animazioni

# Portable Network Graphics (PNG)

- Nato da progetto open source come successore del GIF
- Lossless e lossy
- Diverse tecniche di codifica del colore
  - Indicizzata
  - Livelli di grigio
  - Colore a 24 bit (16 milioni di colori)
- I pixel sono memorizzati in modo da permettere una veloce preview senza aspettare di aver scaricato tutto il file

# Scalable Vector Graphics (SVG)

- Formato vettoriale
- Sintassi XML eventualmente compressa con tecniche usate per file testuali (zip, gzip)
- Standard del World Wide Web Consortium

# Formati per applicazioni mediche

- Health Level 7
- Clinical Document Architecture (CDA)
- CDISC
- DICOM

## Health Level 7 (HL7)

- Formato per scambiare, memorizzare, condividere e recuperare informazioni mediche
- Alla base della fornitura, gestione e valutazione di servizi medici
- Il numero 7 ricorda che HL7 si pone a livello 7 della pila ISO/OSI

# Ambito di impiego di HL7

- Gli ospedali hanno diversi sistemi automatici (hospital information systems)
  - Accettazione pazienti
  - Cartelle cliniche
  - Analisi di laboratorio
- HL7 fornisce interfacce standard, tipi di messaggi e linee guida per far parlare tra loro questi sistemi
- Reference Information Model (RIM) è una parte fondamentale di HL7 e uno standard ISO
  - Esprime il contenuto informativo dei processi ospedalieri
  - Connessioni semantiche e lessicali tra i messaggi previsti da HL7