Modulo di Sistemi Operativi per il corso di Master
RISS: Ricerca e Innovazione nelle Scienze della Salute
Unisa, 17-26 Luglio 2012



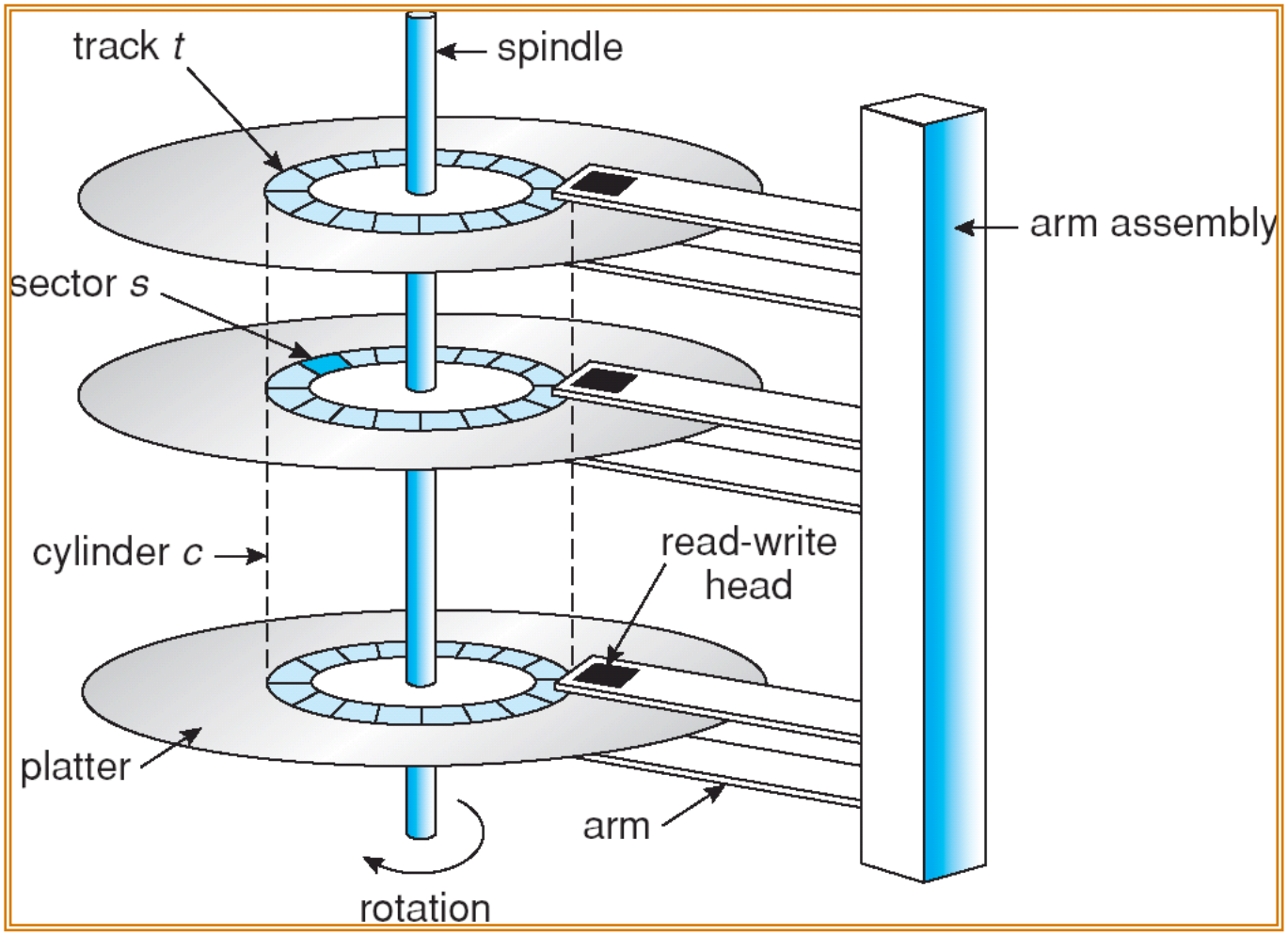
Sistemi Operativi

Gestione della memoria secondaria: dischi, SAN e organizzazioni RAID

Contenuti

- Struttura del disco
- Collegamento dei dischi
- Scheduling del disco
- Gestione del disco
- Strutture RAID

Disco magnetico



Disco magnetico

- I dischi magnetici realizzano la memoria secondaria nei calcolatori moderni
  - I piatti del disco ruotano da 60 a 200 volte al secondo
  - **Velocità di trasferimento:** è la velocità con cui i dati vengono trasferiti dal disco al computer
  - **Tempo di posizionamento (o tempo di accesso):** è il tempo per muovere la testina sul settore desiderato
  - **Caduta della testina:** la testina entra a contatto con la superficie del disco
- I dischi possono essere rimovibili
- Sono collegati al computer tramite **bus di I/O**
  - Più comuni: **EIDE, ATA, SATA, USB, Fibre Channel, SCSI**
  - Un **controller (host controller)** del computer usa il bus per parlare con il **controller (disk controller)** del disco

Nastro magnetico

- Nastro magnetico
  - Usato in passato come dispositivo di memorizzazione secondaria
  - Può contenere grosse quantità di dati
  - Tempo di posizionamento elevato
  - Accesso diretto ~1000 volte più lento del disco
  - Principalmente usato per backup, memorizzazione di dati usati raramente, trasferimenti dati tra sistemi
  - Quando i dati sono sotto la testina, **le velocità di trasferimento sono comparabili** a quelli dei dischi
  - Capacità tipiche vanno da 20 a 200GB
  - Tipologie comuni sono: 4mm, 8mm, 19mm, LTO-2, SDLT

Struttura logica del disco

- Un disco viene visto come un grande array monodimensionale di **blocchi logici**, dove il blocco logico è la più piccola unità di trasferimento (512 byte di solito).
- L'array monodimensionale di blocchi logici è mappato in settori del disco sequenzialmente.
  - Il settore 0 è il primo settore della prima traccia del cilindro più esterno.
  - La mappatura procede in ordine attraverso la traccia, poi attraverso le tracce restanti nel cilindro, e infine attraverso i cilindri restanti, dal più esterno al più interno.

Mapping logico – fisico

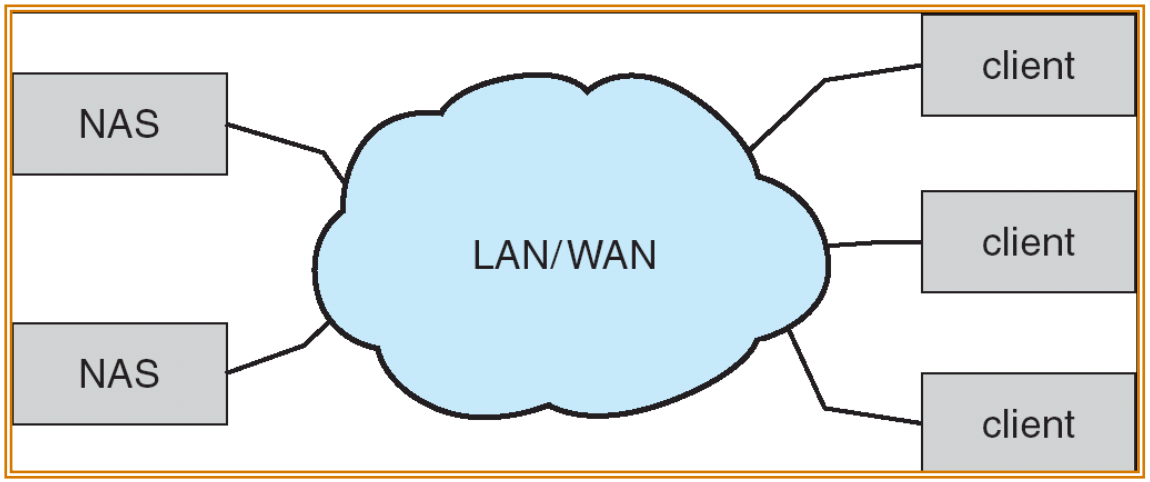
- Alcuni settori possono essere danneggiati
- Il numero di settori per traccia non è costante su alcuni dispositivi
- Dispositivi CVL (constant linear velocity) – la densità di bit per traccia è uniforme
  - tracce esterne contengono più settori
- Dispositivi CAV (constant angular velocity) – la densità di bit decresce dalle tracce interne verso quelle esterne

Collegamento dei dischi

- I dischi connessi direttamente al calcolatore sono accessibili tramite porte di I/O che comunicano con i bus di I/O
- SCSI: supporta fino a 16 dispositivi su un cavo. Uno **SCSI initiator** richiede un'operazione e gli **SCSI target** la realizzano
  - Ciascun target può avere fino a 8 **unità logiche**
- FC è un'architettura seriale ad alta velocità
  - Può avere capacità di commutazione con uno spazio di indirizzamento a 24-bit – la base per **storage area networks** (**SAN**s) in cui molti host sono collegati a molte unità di memorizzazione
  - Variante: a ciclo arbitrato (**arbitrated loop (FC-AL)**) con 126 dispositivi (unità e controller)

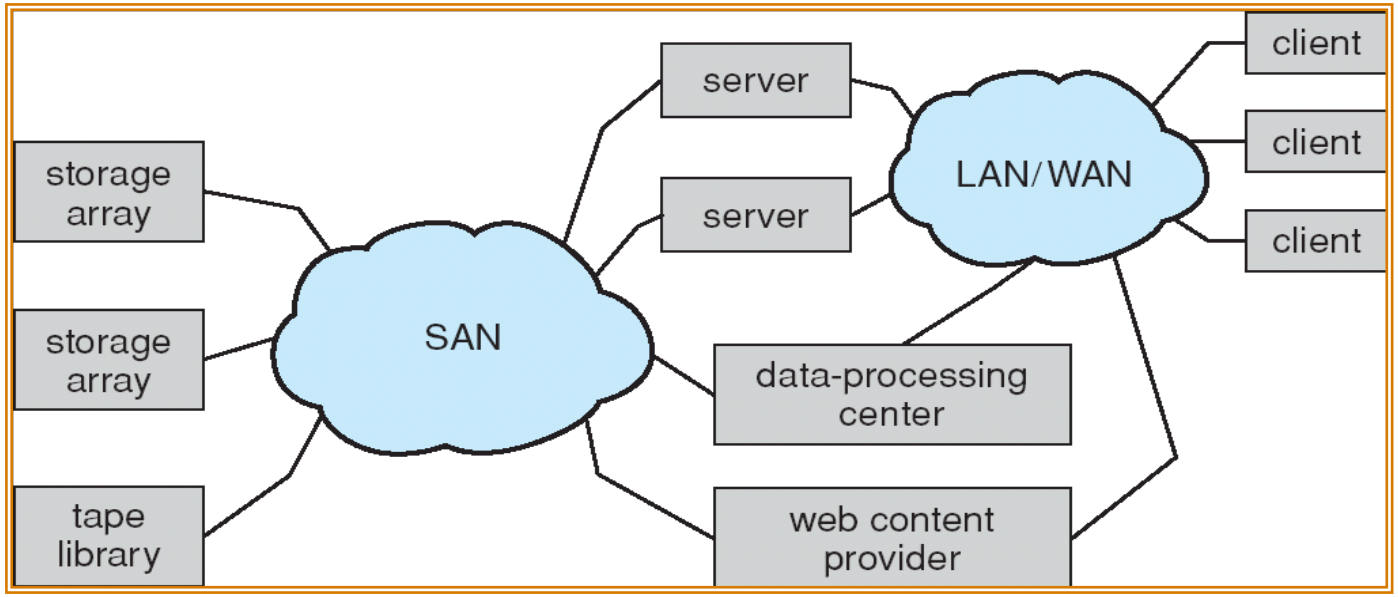
Memorizzazione con collegamento di rete

- La memoria con collegamento di rete (Network-attached storage -**NAS**) è memoria accessibile tramite rete piuttosto che connessioni locali (e.g., bus)
- NFS e CIFS sono protocolli comuni
- Implementata tramite chiamate di procedura remota (RPC) tra l'host e il dispositivo di memorizzazione
- Il protocollo iSCSI usa reti IP per portare su rete il protocollo SCSI



Reti di memoria secondaria

- In grandi ambienti di memorizzazione (e sta diventando sempre più comune)
- Host multipli sono collegati a più array di dischi – schema flessibile



Scheduling del disco

- Uso efficiente dell'hardware per le unità a disco: rapido tempo di posizionamento e ampiezza di banda.
- Il tempo di posizionamento ha due componenti principali:
  - Il **tempo di ricerca** (**seek time**) è il tempo che impiega il braccio del disco a muovere le testine fino al cilindro contenente il settore desiderato.
  - La **latenza di rotazione** (**rotational latency**) è il tempo aggiuntivo speso in attesa che il disco faccia ruotare il settore desiderato sotto la testina.
- Minimizzare il tempo di ricerca.
- Tempo di ricerca ≈ distanza di ricerca.
- L' **ampiezza di banda** (**bandwith**) del disco è data dal numero totale di byte trasferiti diviso per il tempo totale che intercorre fra la richiesta di servizio e il completamento dell'ultimo trasferimento.

Scheduling del disco

- Richiesta di I/O
  - Lettura/scrittura
  - Indirizzo del disco
  - Indirizzo di memoria
  - Numero di byte da trasferire

- Diversi algoritmi esistono per schedulare le richieste di I/O per il disco.

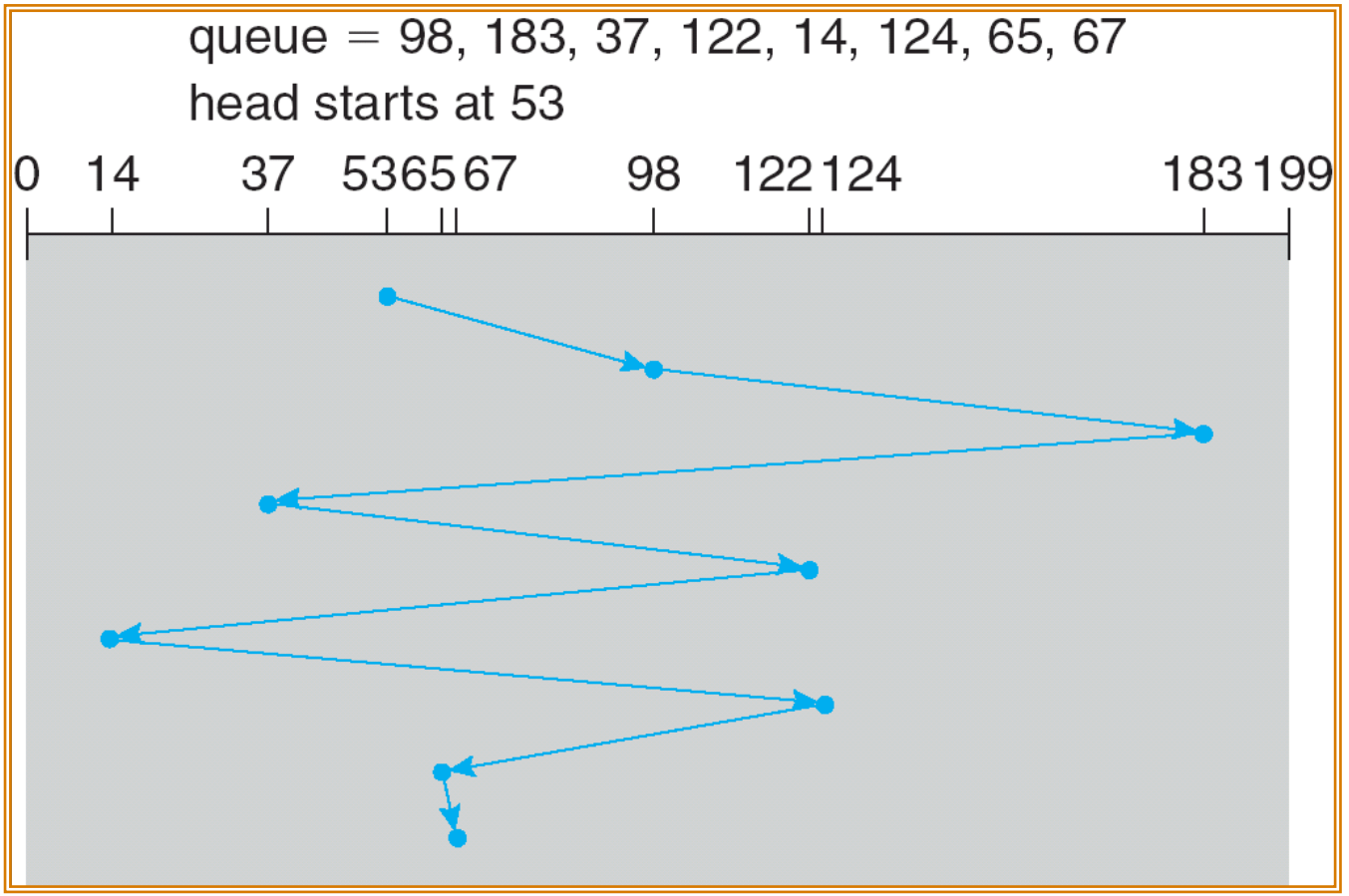
- Li presentiamo usando la seguente coda di richieste (0-199)

  98, 183, 37, 122, 14, 124, 65, 67

  La testina è posizionata sul cilindro 53

FCFS

Il movimento totale della testina è 640 cilindri

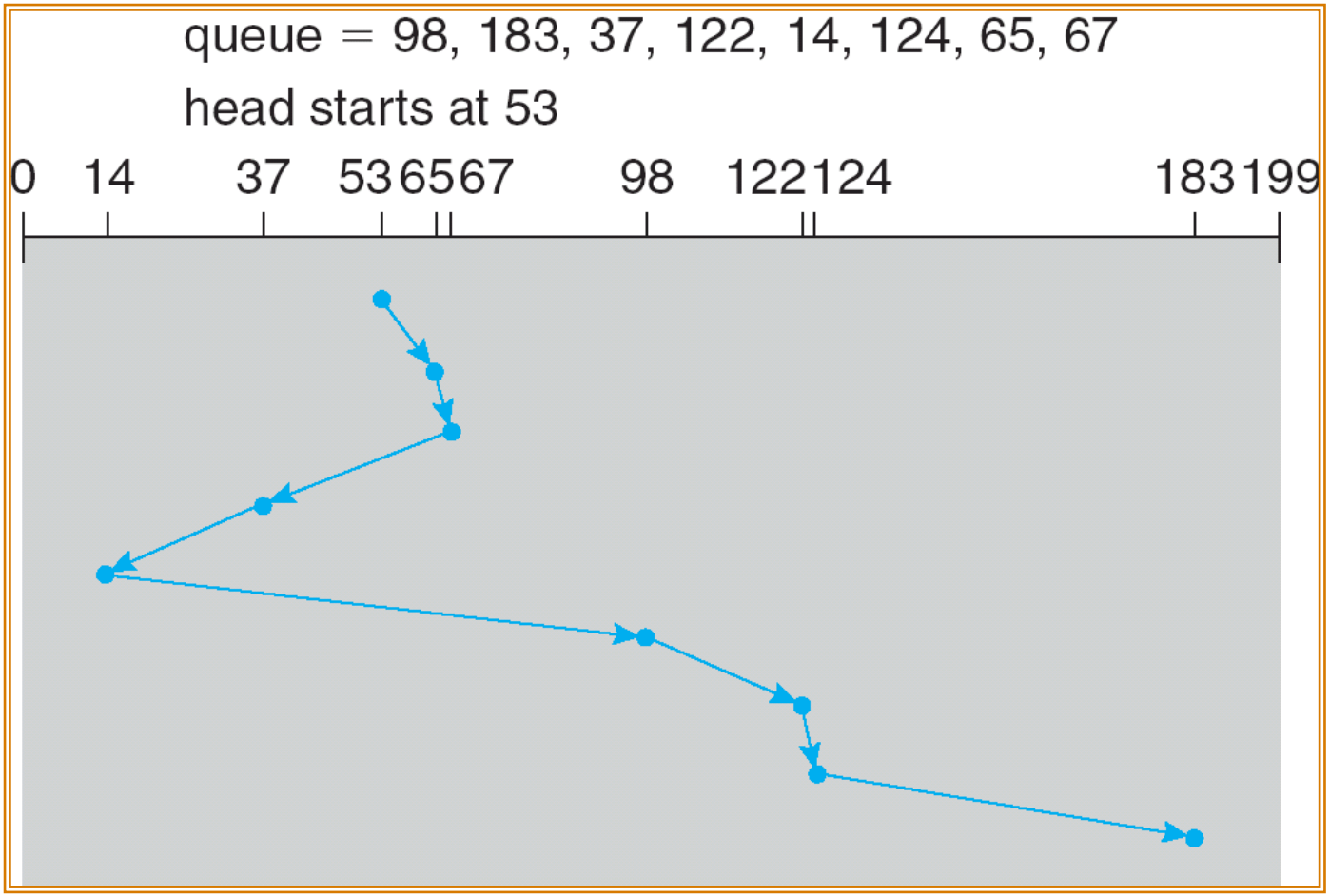


SSTF

- Seleziona la richiesta con il minimo tempo di ricerca dalla posizione corrente della testina.
- La schedulazione SSTF è una forma di schedulazione SJF: può causare attesa indefinita di alcune richieste.
- L'illustrazione mostra che il movimento totale della testina è 236 cilindri.
- Non è ottimale (e.g., nella figura che segue servire prima le richieste per 53, 37 e 14 riduce il movimento totale)

SSTF

Il movimento totale della testina è 263 cilindri

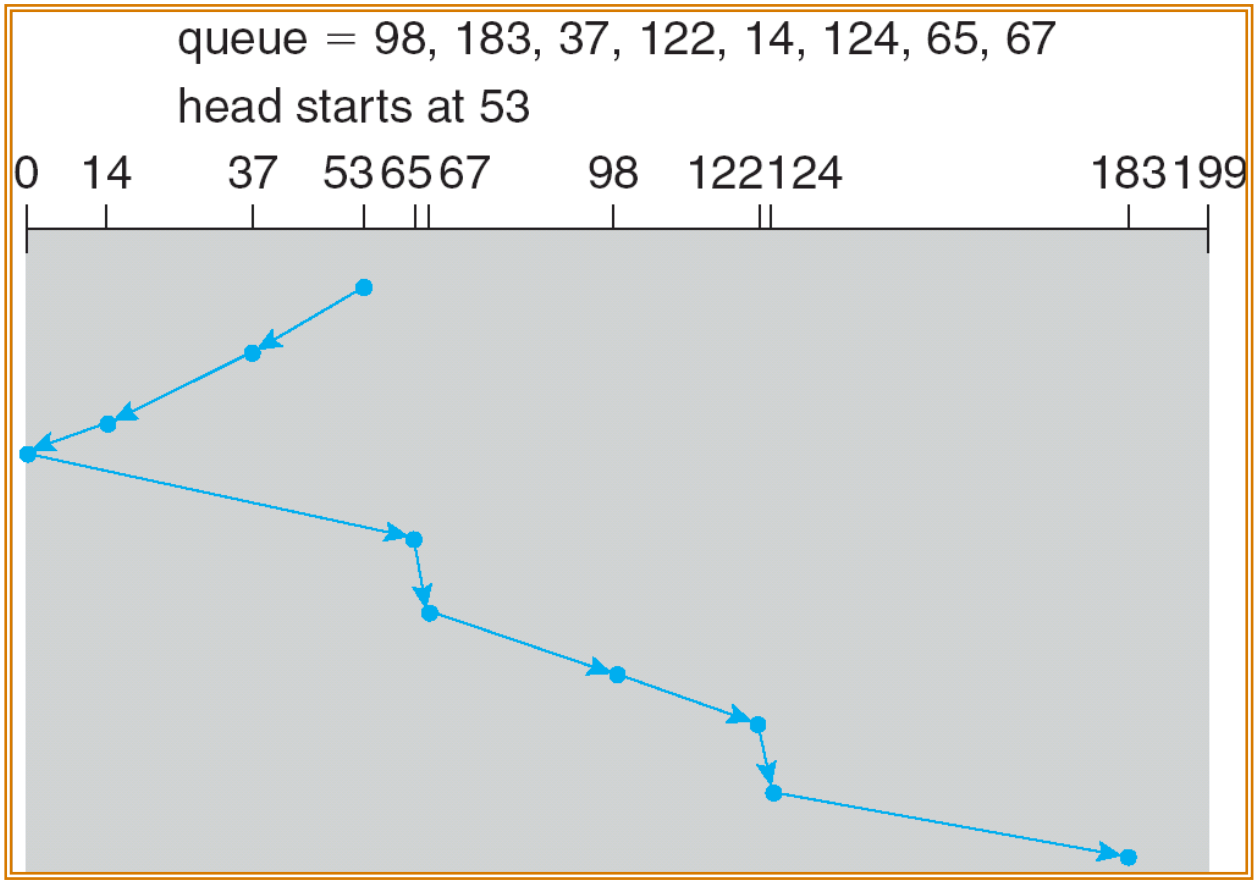


SCAN

- Il braccio del disco si muove da un estremo all'altro estremo, servendo le richieste che incontra. All'altro estremo il movimento viene invertito e il servizio continua.
- Talvolta chiamato **algoritmo dell'ascensore**.
- L'illustrazione mostra che il movimento totale della testina è 208 cilindri.

SCAN

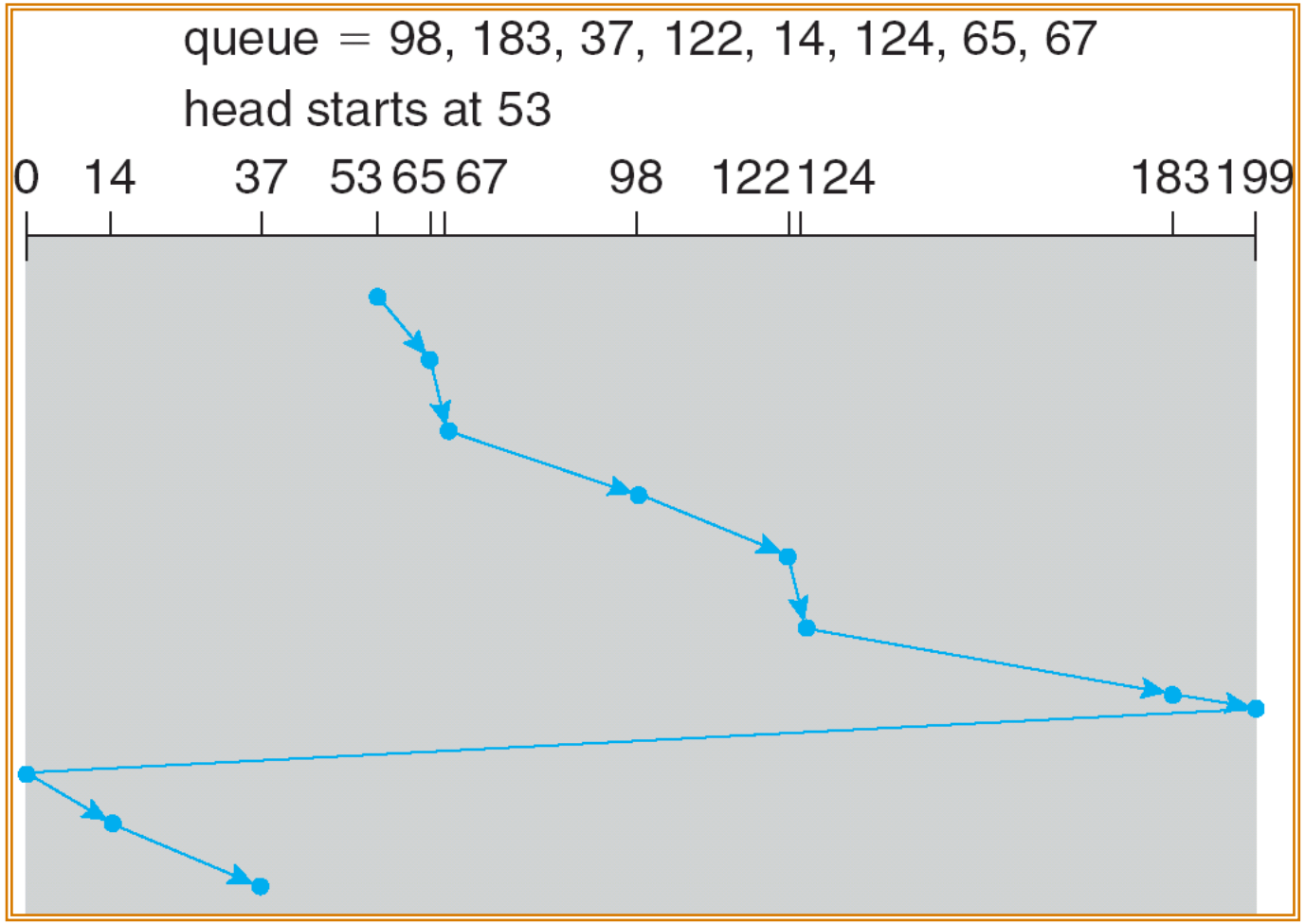
Il movimento totale della testina è 208 cilindri



C-SCAN

- Fornisce un tempo di attesa più uniforme di SCAN.
- Il braccio si muove da un capo all'altro del disco, servendo le richieste lungo il percorso. Quando raggiunge l'altro capo ritorna direttamente all'inizio del disco, senza servire alcuna richiesta durante il ritorno.
- Tratta i cilindri come una lista circolare che si riavvolge dal cilindro finale al primo.

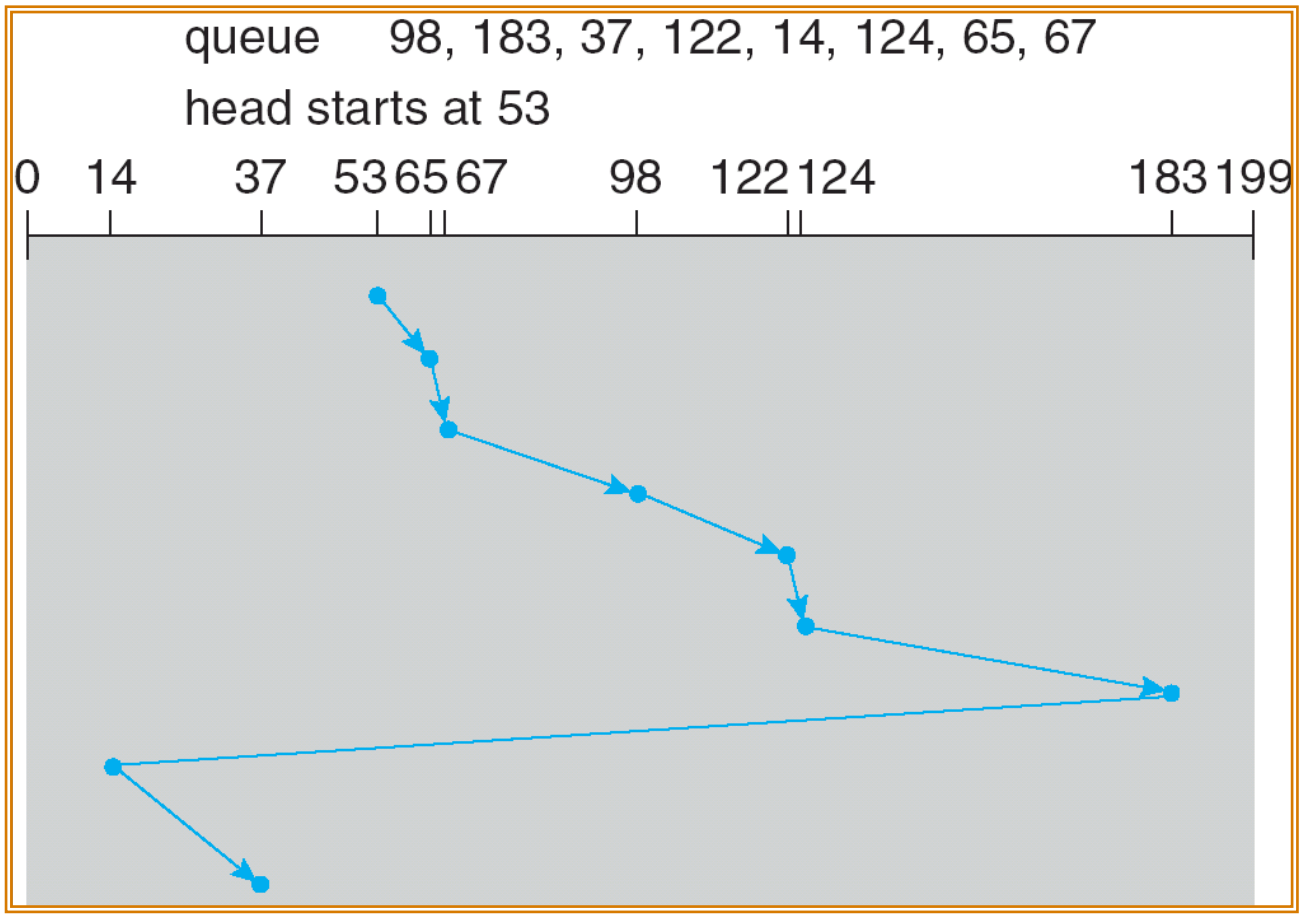
C-SCAN


queue = 98, 183, 37, 122, 14, 124, 65, 67
head starts at 53
0
14
37
53
65
67
98
122
124
183
199

LOOK

- Versione di SCAN e C-SCAN.
- Il braccio arriva fin dove è presente la richiesta finale, per ciascuna delle due direzioni. Lì inverte immediatamente direzione, senza giungere all'estremità del disco.

C-LOOK


queue 98, 183, 37, 122, 14, 124, 65, 67
head starts at 53
0 14 37 53 65 67 98 122 124 183 199

Selezione dell'algoritmo di scheduling del disco

- SSTF è comune e naturale
- SCAN e C-SCAN danno migliori risultati per sistemi che pongono un carico pesante sul disco.
- Le prestazioni dipendono dal numero e dai tipi di richieste.

- Le richieste per servizio su disco possono essere influenzate dal metodo di allocazione dei file (e.g., contigua, concatenata).

- La procedura di schedulazione del disco dovrebbe essere scritta come modulo separato del sistema operativo, in modo da poter essere sostituita con una procedura differente se necessario.

- Sia SSTF che LOOK sono una scelta ragionevole come algoritmo predefinito.

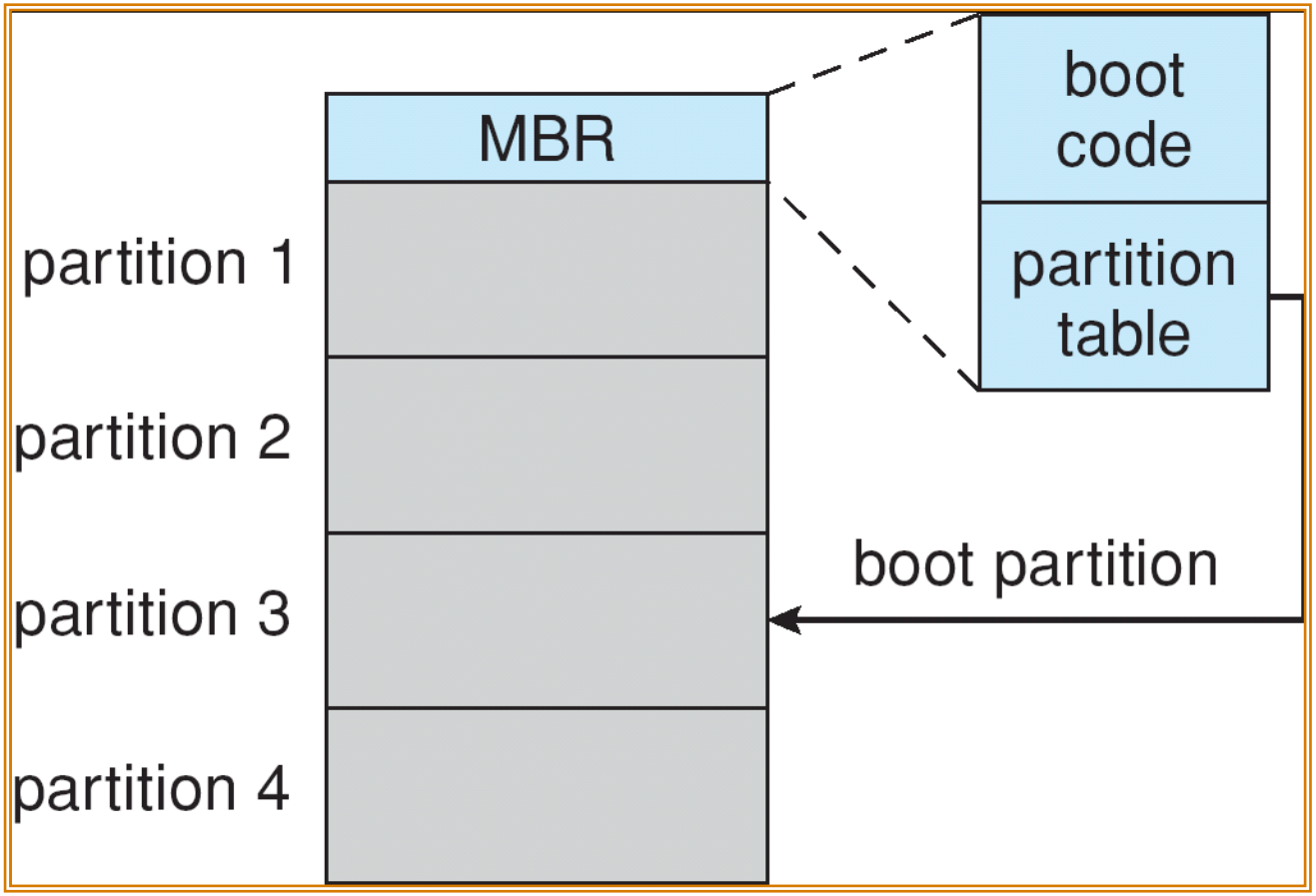
Gestione del disco

- **Formattazione a basso livello**, o **formattazione fisica** — Divide il disco in settori che il controller del disco può leggere e scrivere.
- Settore: header / area dati / trailer (numero settore, ECC)
- Solitamente taglie di 256, 512, o 1024 byte
- Per usare un disco il sistema operativo crea delle partizioni contenenti uno o più cilindri. Ogni partizione è come se fosse un disco separato
- Dopo la creazione delle partizioni, occorre **formattarle logicamente** i.e., il sistema operativo ha bisogno di registrare su disco le proprie strutture dati (creazione di un file system)

Gestione del disco

- Partizione d'avviamento contiene il sistema operativo. Il blocco di avviamento contiene codice per inizializzare il sistema.
  - Il bootstrap loader è immagazzinato nella ROM.
  - Programma d'avviamento completo (blocco d'avviamento).
- Metodi per gestire blocchi difettosi
  - **sector sparing (accantonamento dei settori)**
  - **sector slipping (traslazione dei settori)**

Boot da disco in Windows 2000



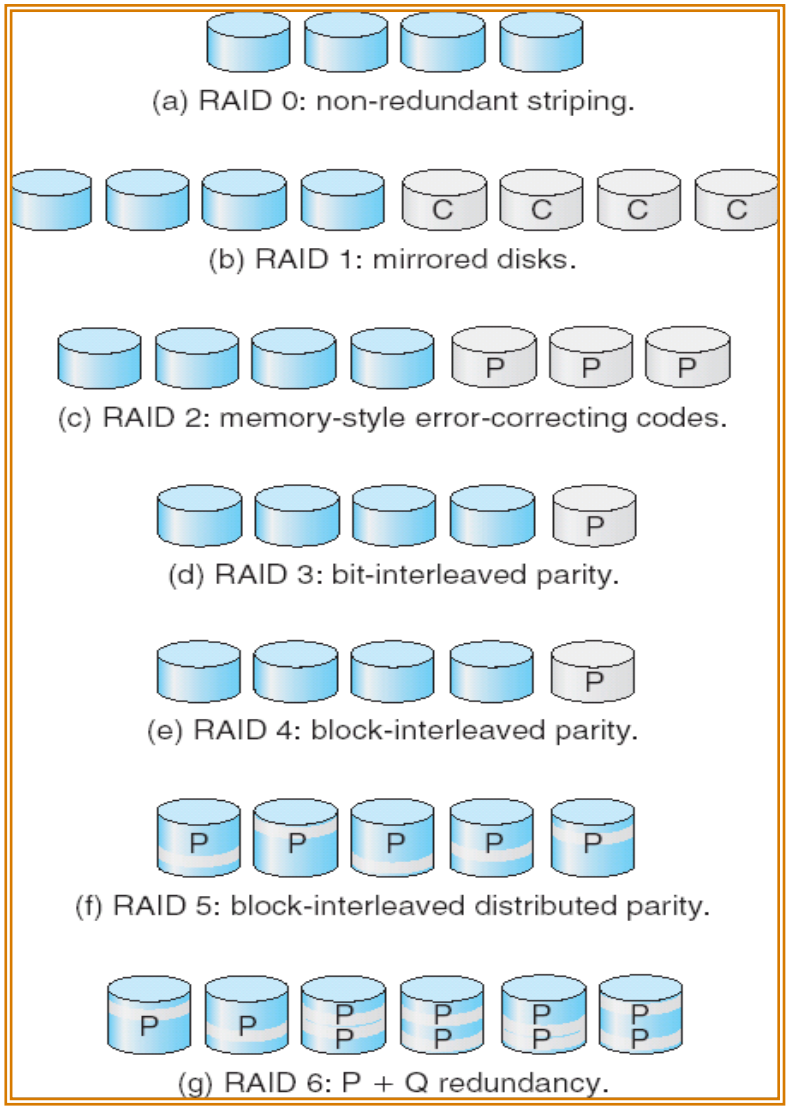
RAID

- **RAID** (redundant arrays of inexpensive disks) – dischi multipli forniscono affidabilità attraverso l'uso di ridondanza.
- Inexpensive / Indipendent
- Esistono sei differenti livelli di organizzazione RAID

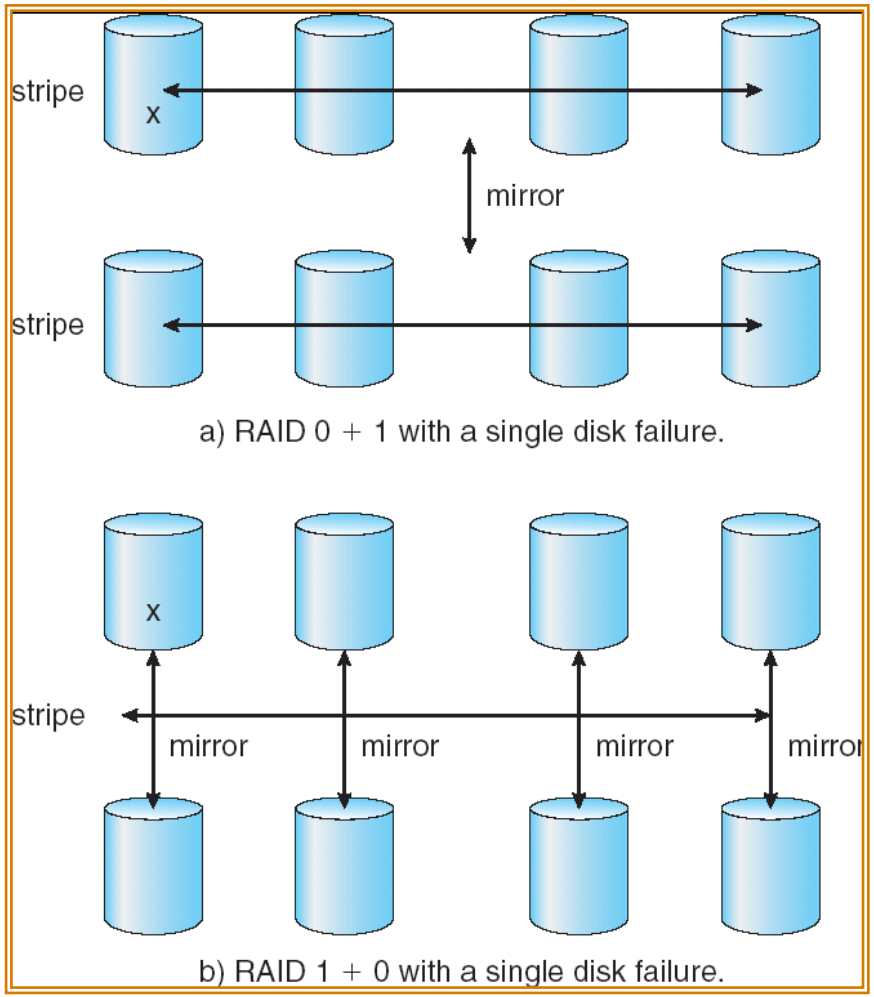
RAID

- Lo spezzettamento dei dischi (**disk striping**) utilizza un gruppo di dischi come una singola unità di memorizzazione.
- Gli schemi RAID aumentano le prestazioni ed aumentano l'affidabilità del sistema di memorizzazione attraverso l'immagazzinamento di dati ridondanti.
  - **Mirroring** o **shadowing** mantengono una copia di ogni disco.

Livelli RAID



RAID (0 + 1) e (1 + 0)


stripe
x
mirror
stripe
a) RAID 0 + 1 with a single disk failure.
x
stripe
mirror
mirror
mirror
mirror
b) RAID 1 + 0 with a single disk failure.

Dispositivi di memorizzazione terziaria

- Il basso costo è la caratteristica che definisce la memorizzazione terziaria.
- Generalmente, la memorizzazione terziaria viene effettuata con **supporti rimovibili**.
- Gli esempi più comuni di supporti rimovibili sono floppy disk, CD-ROM e nastri: sono disponibili anche altri tipi di dispositivi.

Dischi rimovibili

- Floppy disk — sottile disco flessibile ricoperto di materiale magnetico, chiuso in una custodia di plastica protettiva.
  - I dischetti possono contenere circa 1.5 MB; si usa una tecnologia simile per i dischi magnetici rimovibili che contengono fino ad 1 GB.
  - I dischi magnetici rimovibili possono essere veloci quasi quanto i dischi rigidi, anche se sono a più alto rischio di danneggiamento.

Dischi rimovibili

- Un **disco magneto-ottico** registra i dati su un piatto rigido ricoperto di materiale magnetico.
  - Il calore del laser amplifica un campo magnetico grande e debole per memorizzare un bit sul disco.
  - Il raggio laser è anche usato per leggere i dati (Kerr effect).
  - La testina magneto-ottica è più lontana dalla superficie del disco rispetto alla testina di un disco magnetico, e il materiale magnetico è ricoperto da uno strato protettivo di plastica o vetro. Più resistente alla caduta delle testine.
- I **dischi ottici** non usano il magnetismo, bensì materiali speciali che possono essere alterati dal raggio laser per avere punti relativamente scuri o luminosi.

Dischi WORM

- I dati su dischi con lettura-scrittura possono essere modificati più e più volte.
- I dischi WORM ("Write Once, Read Many Times") possono essere scritti solamente una volta.
- Sottile pellicola di alluminio interposta tra due dischi di vetro o plastica.
- Per scrivere un bit, il driver usa un raggio laser per creare un piccolo buco attraverso l'alluminio. Le informazioni possono essere distrutte ma non alterate.
- Molto durevole ed affidabile.
- I **dischi a sola lettura** (**read-only**), quali CD-ROM e DVD, provengono dalla fabbrica con dati preregistrati.

Nastri

- Rispetto ad un disco, un nastro è meno costoso e contiene più dati, ma l'accesso casuale è molto più lento.
- Il nastro è un mezzo economico per usi che non richiedono un accesso casuale veloce, e.g., tenere copie di riserva dei dati del disco, memorizzare grossi volumi di dati.
- Le grandi installazioni che usano nastri, utilizzano tipicamente dei robot per cambiare i nastri, che li spostano fra le unità e gli spazi di archiviazione in una libreria di nastri.
  - stacker – libreria che immagazzina alcuni nastri.
  - silo – libreria che immagazzina migliaia di nastri.
- Un file residente su disco, che non sarà più necessario per un po' di tempo, può venire **archiviato** su nastro; il computer potrà **ricaricarlo sul** disco per un uso attivo.

Gestione della memorizzazione gerarchica (HSM)

- Un sistema di memorizzazione gerarchica estende la gerarchia di memorizzazione oltre la memoria centrale e la memoria secondaria per incorporare la memorizzazione terziaria. La memorizzazione terziaria è solitamente implementata come un jukebox di nastri o di dischi rimovibili.
- Il modo usuale per incorporare la memorizzazione terziaria è di estendere il file system.
  - I file piccoli, e usati frequentemente, rimangono sul disco magnetico.
  - I file grandi e vecchi che non sono più usati attivamente vengono archiviati nel jukebox.
- L'HSM si trova solitamente nei centri di calcolo con supercomputer e in altre grandi installazioni che hanno quantità di dati enormi.

Velocità

- La velocità della memorizzazione terziaria ha due aspetti: la larghezza di banda e la latenza.

- La larghezza di banda si misura in byte al secondo.
  - **Larghezza di banda sostenuta** (**sustained bandwith**) – velocità media di trasferimento, i.e., numero di byte/ tempo di trasferimento.
  - **Larghezza di banda effettiva** (**effective bandwith**) – calcola la media sull'intero tempo di I/O, compreso il tempo per seek o locate e il tempo per lo scambio delle cartucce in un jukebox.

Velocità

- Latenza di accesso – tempo necessario per localizzare i dati.
  - Tempo di accesso per un disco – sposta il braccio verso il cilindro selezionato e aspetta la latenza rotazionale; < 5 millisecondi.
  - Un accesso casuale al nastro richiede di avvolgere il nastro sulle bobine finché il blocco selezionato non raggiunge la testina del nastro; può richiedere decine o centinaia di secondi.
  - In generale l'accesso casuale all'interno di un nastro è oltre mille volte più lento dell'accesso casuale su disco.
- Il basso costo di memorizzazione terziaria è legato alla disponibilità di molte cartucce poco costose che condividono alcuni lettori costosi.

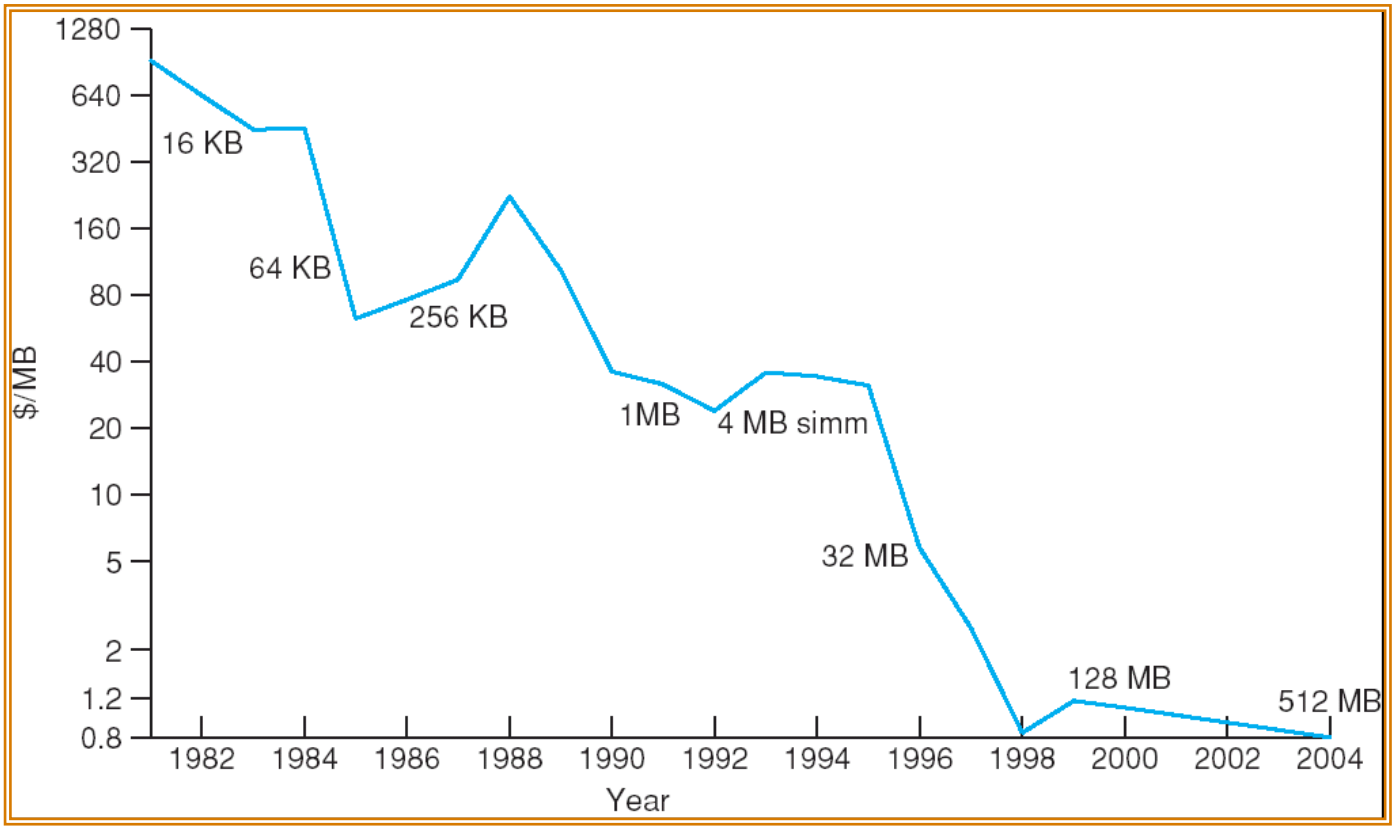
Affidabilità

- I dischi magnetici fissi sono più affidabili dei dischi rimovibili e dei nastri.
- I dischi ottici sono più affidabili dei dischi magnetici e dei nastri
- Un guasto alla testina in un disco rigido, in genere, distrugge i dati, mentre un guasto di un'unità a nastro o dell'unità del disco ottico spesso lascia la cartuccia dei dati intatta.

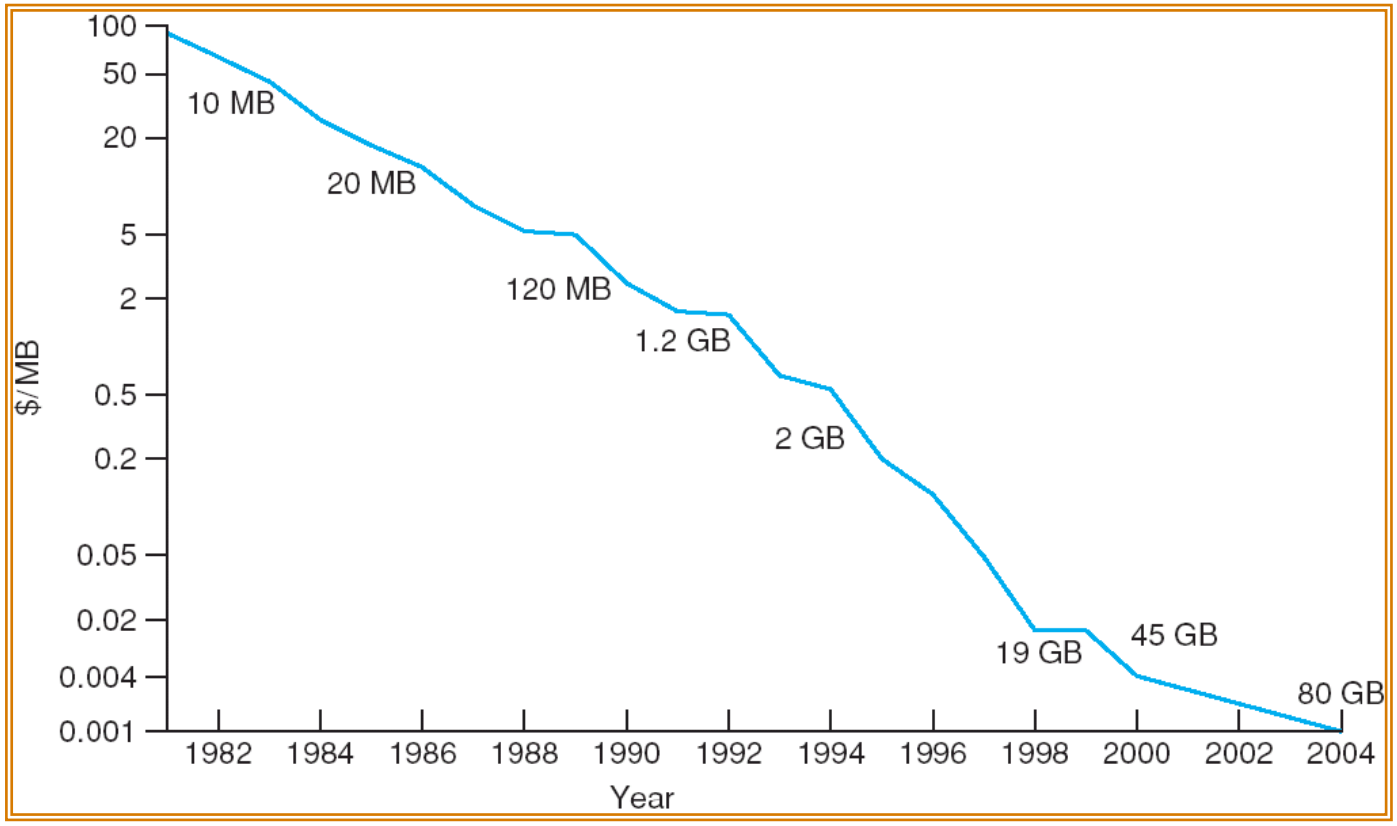
Costo

- La memoria principale è molto più costosa di un disco.
- Il costo per megabyte dell'hard disk è competitivo con il nastro magnetico se un drive supporta un solo nastro.
- Il costo globale di memorizzazione su nastro diventa conveniente quando il numero di nastri è considerevolmente maggiore del numero di drive che le gestiscono.
- Il prezzo al megabyte dei dischi magnetici si è ridotto di oltre 4 ordini di grandezza dal 1981 al 2004. Quello della memoria centrale di 3 ordini.

Prezzo per Megabyte: DRAM, dal 1981 al 2004



Prezzo per Megabyte: dischi magnetici, dal 1981 al 2004



Prezzo per Megabyte: nastro magnetico, dal 1984 al 2000

