Sistemi Operativi A
Parte VI - La memoria secondaria

Augusto Celentano
Università Ca' Foscari Venezia
Corso di Laurea in Informatica

Dischi magnetici

- Proprietà principali e parametri
  - Velocità di rotazione più comuni: 4200, 5400, 7200, 10000 giri/minuto
  - Velocità di trasferimento: istantanea e a regime
  - Tempo di posizionamento: comprende il tempo necessario per muovere il braccio portatestina sul cilindro richiesto (seek time) e il tempo di rotazione necessario per portare il settore richiesto sotto la testina di lettura (rotational latency)
  - Fissi o rimovibili, a sola lettura o R/W
- Sono collegati attraverso un'interfaccia di I/O
  - Diverse tecnologie: ATA/IDE, EIDE, Serial ATA, USB, FW, SCSI
  - La gestione è curata da un controller logicamente diviso in due parti: verso l'host e verso l'unità a disco

Nastri magnetici

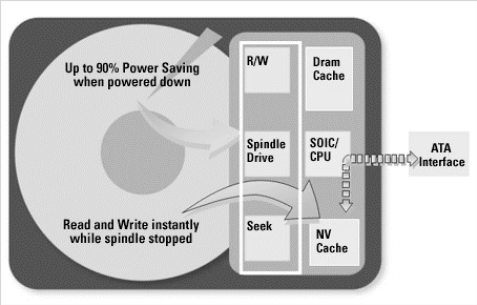
- Proprietà principali
  - Sono stati i primi supporti per la memorizzazione esterna di dati
  - Grande capacità, lunghi tempi di conservazione (~)
  - Accesso sequenziale, bassa velocità di posizionamento, velocità di trasferimento accettabile
  - Usati per copie di archivio e per trasferimento dati tra sistemi
  - Dimensioni tipiche 20-200GB per nastro
  - Costo relativo molto basso, poco pratici
  - Cartucce, sistemi robotizzati

Dischi ottici

- CD-DVD
  - Nati per applicazioni audio-video, accesso prevalentemente sequenziale
  - Tecnologia di lettura e scrittura laser, capacità variabile in funzione della lunghezza d'onda del laser
  - Costo molto basso, velocità elevata in caso di accesso sequenziale
  - Durata abbastanza elevata (20-100 anni?)
  - Utilizzo prevalente per distribuzione di informazioni e archivio personale
  - Supporti non riutilizzabili (eccetto dispositivi RW)

Memorie ibride

- Disco magnetico + memoria cache non volatile



Struttura logica dei dischi magnetici

- I dischi sono considerati un grande vettore monodimensionale di *blocchi logici*, dove un blocco logico è la minima unità di trasferimento.
- Il vettore corrisponde in modo sequenziale ai settori del disco:
  - Il settore 0 è il primo settore della prima traccia sul cilindro più esterno
  - La corrispondenza prosegue ordinatamente lungo la prima traccia, quindi lungo le rimanenti tracce del primo cilindro, e così via di cilindro in cilindro, dall'esterno verso l'interno
  - Eccezione: tracce di riserva

Scheduling del disco

- Il sistema operativo è responsabile di una gestione efficiente delle risorse fisiche
  - tempi d'accesso contenuti e ampiezze di banda elevate
- Il tempo d'accesso ha due componenti principali:
  - il tempo di ricerca (*seek time*) è il tempo necessario affinché il braccio dell'unità a disco sposti le testine fino al cilindro contenente il settore desiderato
  - la latenza di rotazione (*rotational latency*) è il tempo aggiuntivo necessario perché il disco ruoti finché il settore desiderato si trovi sotto la testina
  - minimizzare il tempo d'accesso ~= minimizzare la distanza percorsa
- L'ampiezza di banda del disco (*disk bandwidth*) è il numero totale di byte trasferiti diviso il tempo totale intercorso fra la prima richiesta e il completamento dell'ultimo trasferimento

Calcolo del tempo di accesso

- Il seek time dipende dalla distanza tra le tracce

$$S \cong a + bN$$

~ 1 mS fra tracce adiacenti
~ 100 mS per l'intero disco (~10^4 tracce)

- La latenza di rotazione dipende dalla velocità di rotazione del disco

$$L_{mS} = 1/2 \times 60/v_{rpm} \times 1000$$

4200 rpm	7,14mS
5400 rpm	5,56mS
7200 rpm	4,17mS
10000 rpm	3 mS

- Il seek time è dominante

Scheduling del disco

- Esistono numerosi algoritmi di scheduling
  - First Come First Served (FCFS)
  - Shortest Scan Time First (SSTF)
  - Sequential Scan (SCAN, C-SCAN, LOOK, C-LOOK)
- Consideriamo, ad esempio, una coda di richieste nell'ordine seguente (0-199).

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

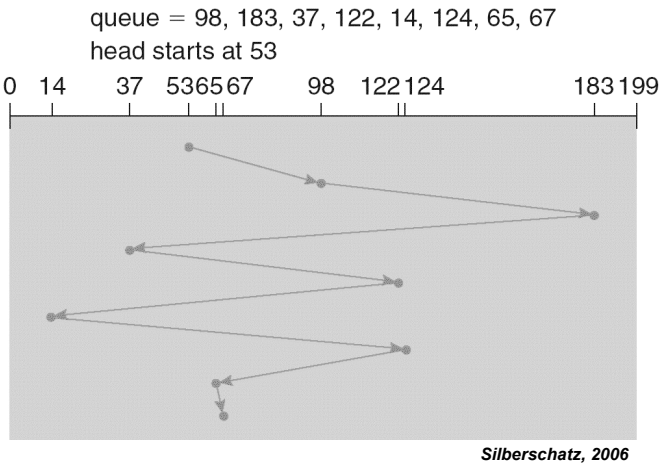
  Puntatore inizialmente al cilindro 53

Scheduling FCFS

- Seleziona le richieste nell'ordine di arrivo.
- Lo scheduling FCFS è una forma di scheduling semplice e *fair*
- Può causare lunghi tempi di posizionamento se le richieste che arrivano riguardano aree molto distanti del disco
  - avvicendamento di processi diversi

FCFS

Movimento totale: 640 cilindri



Silberschatz, 2006

Scheduling SSTF

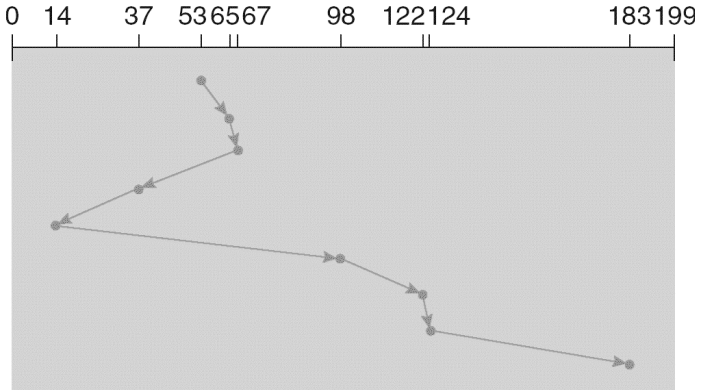
- Seleziona la richiesta con il minor tempo di ricerca rispetto all'attuale posizione della testina.
- Lo scheduling SSTF è essenzialmente una forma di scheduling per brevità (come SJF, *shortest job first*) e, al pari di questo, può condurre a situazioni di attesa indefinita (*starvation*) di alcune richieste
  - es. un insieme di processi prioritari che leggono in una zona limitata del disco possono bloccare un processo che legge in una zona distante
  - soluzione: due code di richieste, mentre si serve l'una le richieste sono accodate sull'altra

SSTF

Movimento totale: 236 cilindri

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

head starts at 53



Silberschatz, 2006

Scheduling per scansione (SCAN)

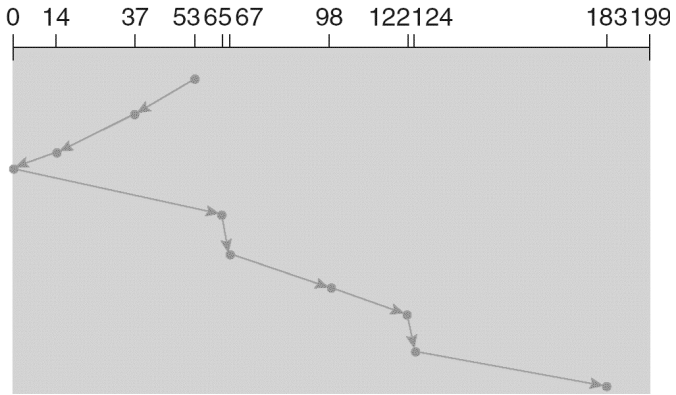
- Secondo l'algoritmo SCAN il braccio dell'unità a disco parte da un estremo del disco e si sposta verso l'altro estremo
  - serve le richieste mentre attraversa i cilindri, fino a che non giunge all'altro estremo del disco
  - il braccio inverte la marcia e la procedura continua nel verso opposto
- L'algoritmo SCAN è chiamato anche *algoritmo dell'ascensore*
  - il braccio dell'unità a disco si comporta come un ascensore che serve prima tutte le richieste in salita e poi tutte quelle in discesa

SCAN

Movimento totale: 236 cilindri

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

head starts at 53



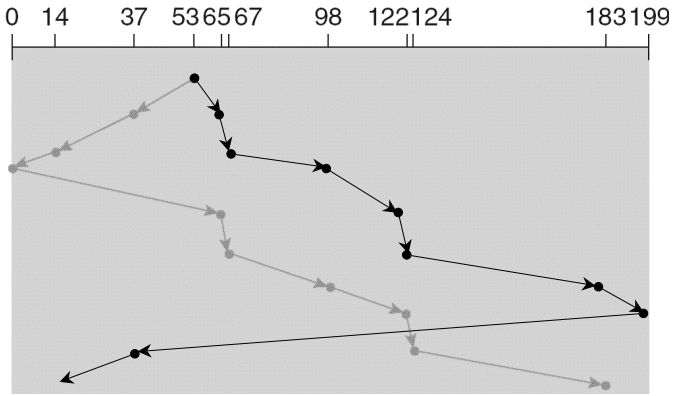
Silberschatz, 2006

SCAN

Movimento totale: 331 cilindri

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

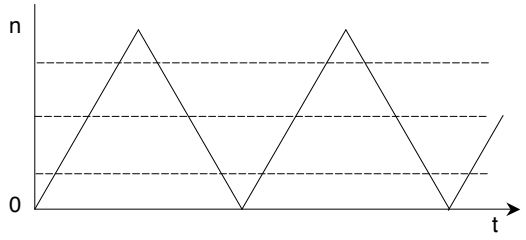
head starts at 53



Silberschatz, 2006

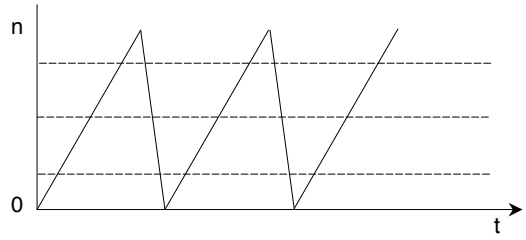
SCAN - problema

- L'algoritmo SCAN non garantisce tempi d'attesa uniformi
  - I cilindri sulle tracce esterne sono attraversati con frequenza variabile rispetto ai cilindri centrali



Scheduling per scansione circolare (C-SCAN)

- L'algoritmo C-SCAN tratta il disco come una lista circolare, cioè come se il primo e l'ultimo cilindro fossero adiacenti
  - come lo SCAN, sposta la testina da un estremo all'altro del disco, servendo le richieste lungo il percorso
  - quando la testina giunge all' estremo del disco ritorna all'inizio senza servire richieste durante il viaggio di ritorno
  - garantisce tempi di attesa più uniformi

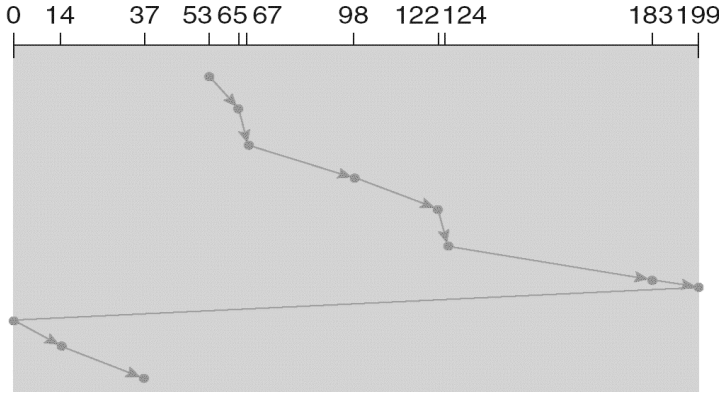


C-SCAN

Movimento totale: 382 cilindri

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

head starts at 53

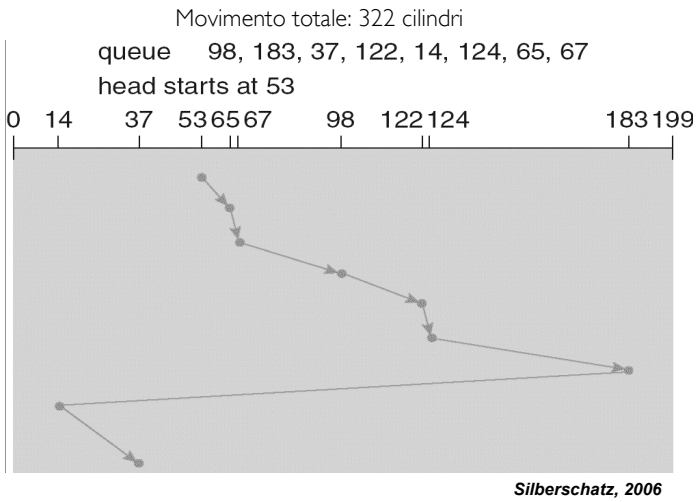


Silberschatz, 2006

LOOK, C-LOOK

- Varianti di SCAN, C-SCAN
- Il braccio si sposta solo finché ci sono altre richieste da servire in ciascuna direzione, dopo di che cambia immediatamente direzione, senza giungere all'estremo del disco

C-LOOK



Silberschatz, 2006

Scelta di un algoritmo di scheduling

- Le prestazioni dipendono in larga misura dal numero e dal tipo di richieste
- Le richieste di I/O per l'unità a disco possono essere notevolmente influenzate dal metodo adottato per l'allocazione dei file
- SSTF è molto comune e naturalmente semplice
- SCAN e C-SCAN offrono migliori prestazioni in sistemi che sfruttano molto le unità a disco
- Sia SSTF sia LOOK costituiscono un ragionevole algoritmo di partenza L'algoritmo di scheduling del disco dovrebbe costituire un modulo a sé stante del sistema operativo così da poter essere sostituito da un altro algoritmo qualora ciò fosse necessario

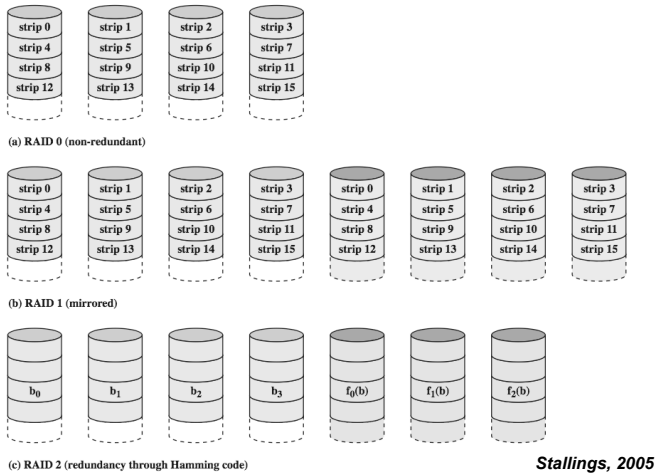
Strutture RAID

- Le **batterie ridondanti di dischi** (RAID, *redundant array of independent/inexpensive disk*) hanno lo scopo di affrontare i problemi di prestazioni e affidabilità.
- La tecnica RAID è organizzata su diversi livelli (0-6)

Strutture RAID

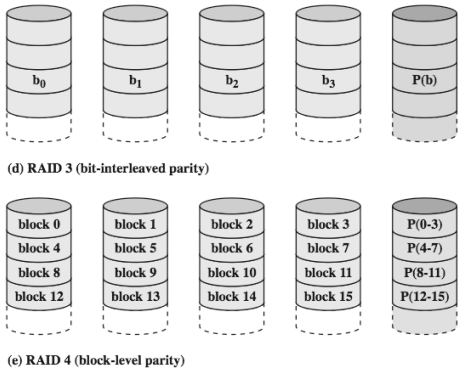
- L'evoluzione tecnologica ha reso le unità a disco progressivamente più piccole e meno costose tanto che oggi è possibile, senza eccessivi sforzi economici, equipaggiare un sistema di calcolo con molti dischi
- La presenza di più dischi, qualora si possano usare in parallelo, rende possibile l'aumento della frequenza alla quale i dati si possono leggere o scrivere
- Gli schemi RAID migliorano l'affidabilità della memoria secondaria poiché diventa possibile memorizzare le informazioni in più dischi in modo ridondante
  - La copiatura speculare *(mirroring* o *shadowing*) mantiene un duplicato di ciascun disco
  - L'organizzazione con blocchi intercalati a parità distribuita utilizza meno la ridondanza

Livelli RAID (0-2)



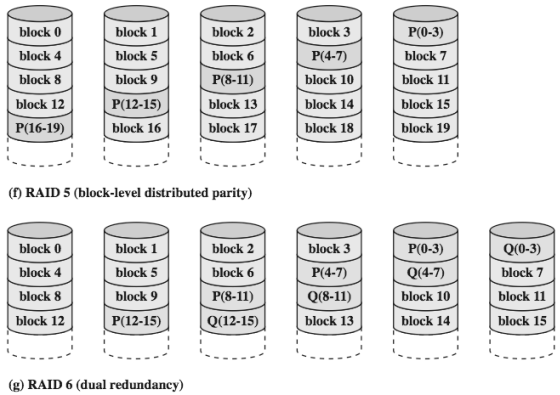
Stallings, 2005

Livelli RAID (3-4)



Stallings, 2005

Livelli RAID (5-6)



Stallings, 2005

RAID

- Performance
  - il data path dai dischi alla memoria (controller, bus, ecc) deve essere in grado di sostenere le maggiori performance
  - l'obiettivo è quello di
    - Ridurre il tempo di accesso per accessi a grandi quantità di dati
    - Ridurre il tempo di risposta (tramite bilanciamento del carico) per accessi multipli indipendenti a piccole porzioni di dati
- Affidabilità
  - la presenza di più dischi aumenta le probabilità di guasto
  - per compensare questa riduzione di affidabilità, RAID utilizza la ridondanza nella memorizzazione (*mirroring*, meccanismi di parità)

RAID 0 (striping)

- Descrizione
  - il sistema RAID viene visto come un disco logico
  - i dati nel disco logico vengono suddivisi in strip (e.g., settori, blocchi, byte, bit oppure altro)
  - strip consecutive sono distribuiti su dischi diversi, aumentando le performance nell'accesso ai dati

RAID 0 (striping)

- Vantaggi
  - più richieste possono essere servite in parallelo
  - stripes a livello di blocco: se due richieste di I/O riguardano blocchi indipendenti di dati, è possibile che i blocchi siano su dischi differenti
  - stripes a livello < blocco: una richiesta di un blocco viene servita in tempo minore (blocco più grande nello stesso tempo)
- Ma...
  - Non è un membro "a tutti gli effetti" della famiglia RAID, perchè non utilizza meccanismi di ridondanza
  - Può essere utilizzato per applicazioni in cui l'affidabilità non è un grosso problema, ma lo sono la velocità e il basso costo

RAID 1 (mirroring)

- Descrizione
  - Adotta uno stile di ridondanza semplice: *mirroring* (*shadowing*)
  - I dati di ogni disco sono copiati in modo speculare su un altro disco di un secondo insieme
  - Come prima, il sistema è basato su striping, ma questa volta ogni strip viene mappato su due dischi diversi

RAID 1

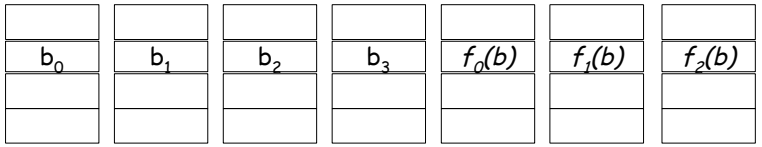
- Performance
  - ogni richiesta di lettura può essere servita da uno qualsiasi dei due dischi che ospitano il dato
    - si può scegliere quello con tempo di seek minore
  - una richiesta di scrittura deve essere portata a termine su ambedue i dischi
    - è legato alla più lenta delle due scritture
- Ridondanza
  - il recovery è molto semplice
    - se un disco si guasta, i dati possono essere recuperati dalla sua copia speculare
    - è quindi necessario sostituire il disco con la copia
- Il costo per unità di memorizzazione raddoppia

RAID 2-3 (accesso parallelo)

- Accesso parallelo
  - tutti i dischi partecipano all'esecuzione di ogni richiesta di I/O
  - i dischi sono sincronizzati in modo che le testine di lettura siano nella stessa posizione allo stesso istante
  - suddivisione fra dischi dati e dischi parità
    - un codice di correzione di errore o di parità viene calcolato a partire dai bit corrispondenti dei dischi dati
    - questo codice viene memorizzato nei dischi parità
  - si utilizza *data striping*, con stripes molto piccoli (bit, byte, word)

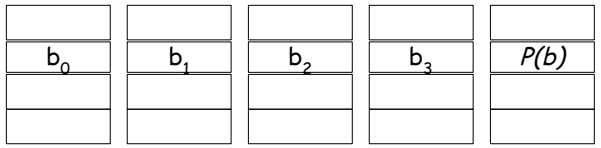
RAID 2

- Descrizione
  - ECC (error correction code) è basato tipicamente sul codice di Hamming, con distanza 3
    - permette di correggere errori fino a un bit (e di rilevare errori fino a due bit)
  - il numero di dischi di parità è proporzionale al logaritmo del numero di dischi di dati
  - è costoso



RAID 3

- Descrizione
  - il codice calcolato è un semplice bit di parità
  - meno costoso: è richiesto un solo disco di parità
  - idea:
    - i dischi hanno già dei meccanismi interni di controllo degli errori
    - una lettura errata viene segnalata dal disco interessato
    - il bit di parità consente quindi di correggere l'errore



Dispositivi per la memorizzazione terziaria

- La memoria terziaria è destinata a funzioni di archivio permanente e non continuamente on-line
  - basso costo per unità di memorizzazione
  - mezzi rimovibili (rete?)