Dischi

caratteristiche
e
prestazioni

06/03/07

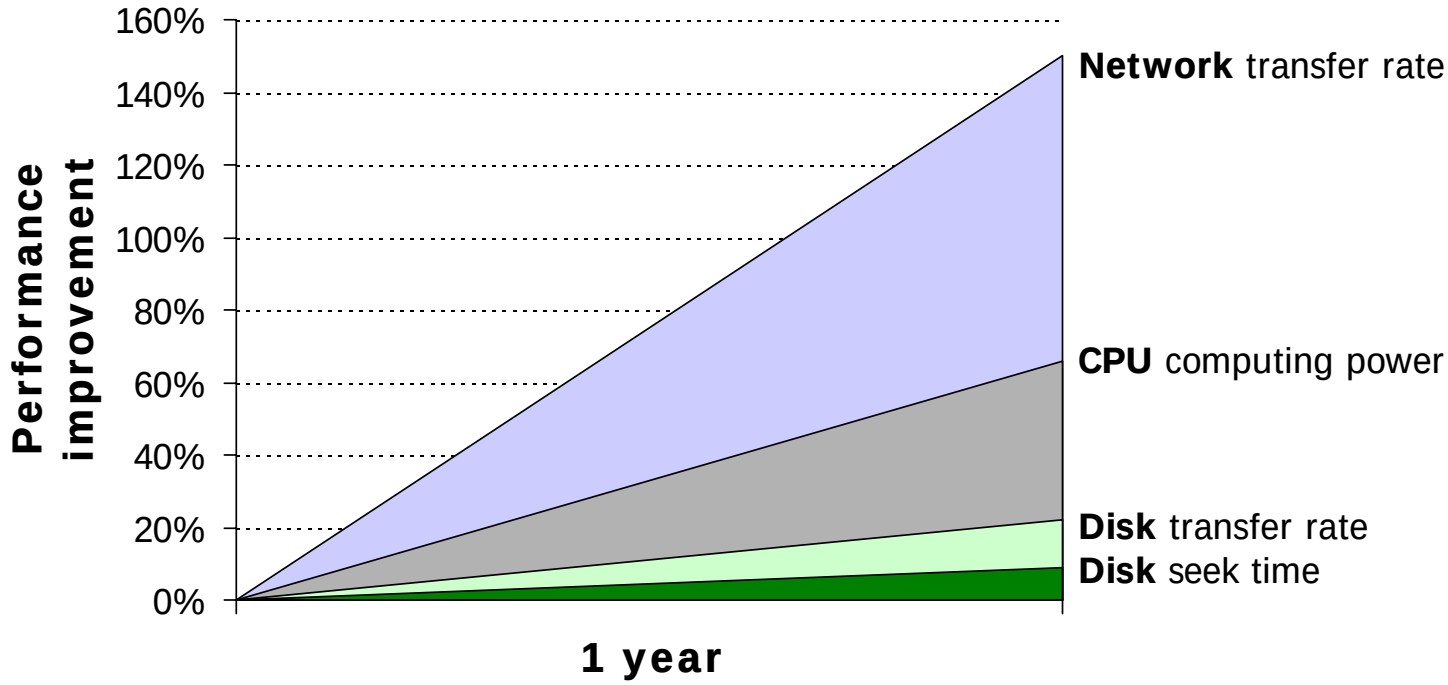
indice

- dischi
  - caratteristiche generali
  - prestazioni
  - tecniche per migliorare le prestazioni di I/O

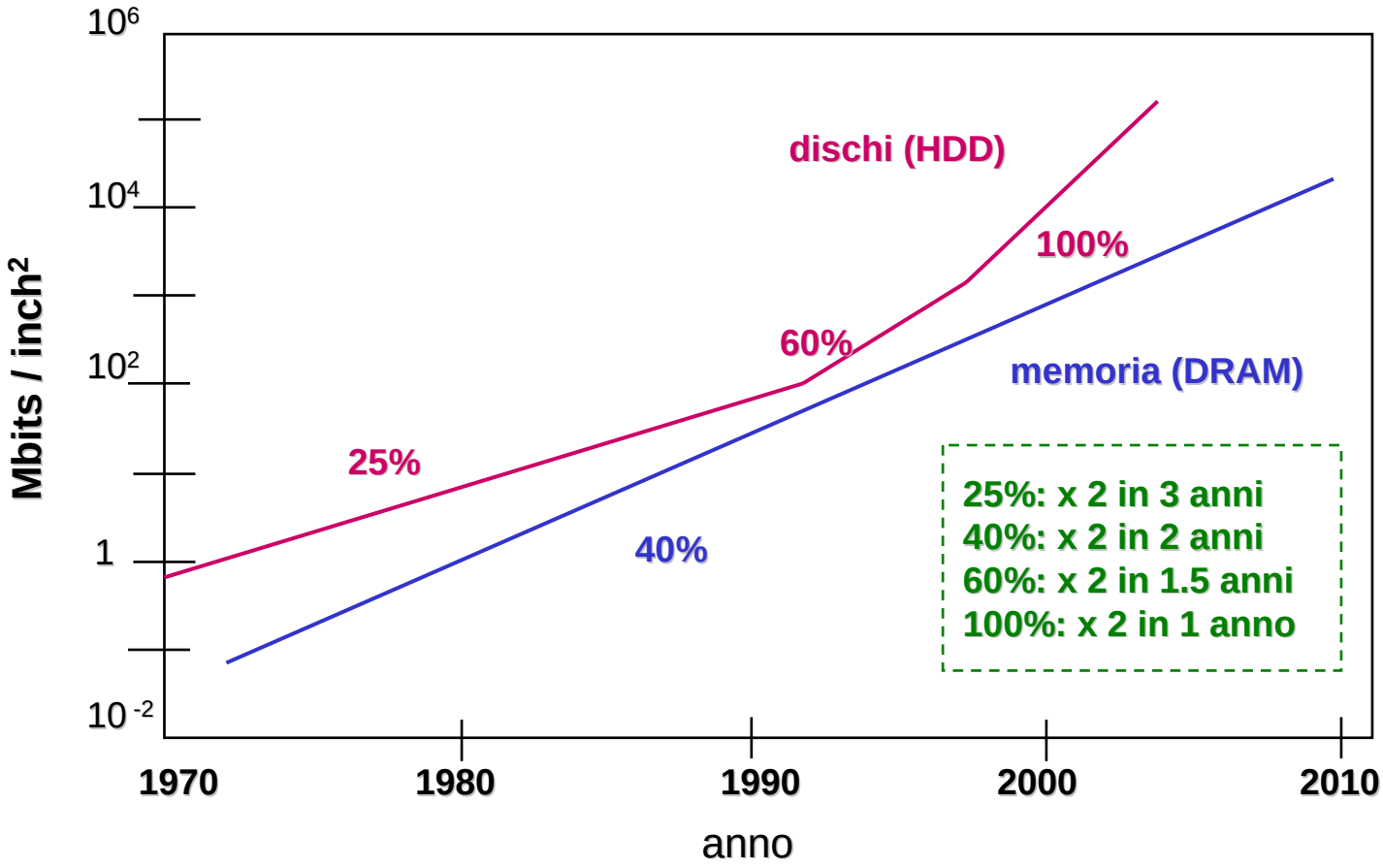
dischi:
caratteristiche generali

evoluzione delle capacità (legge di Moore)

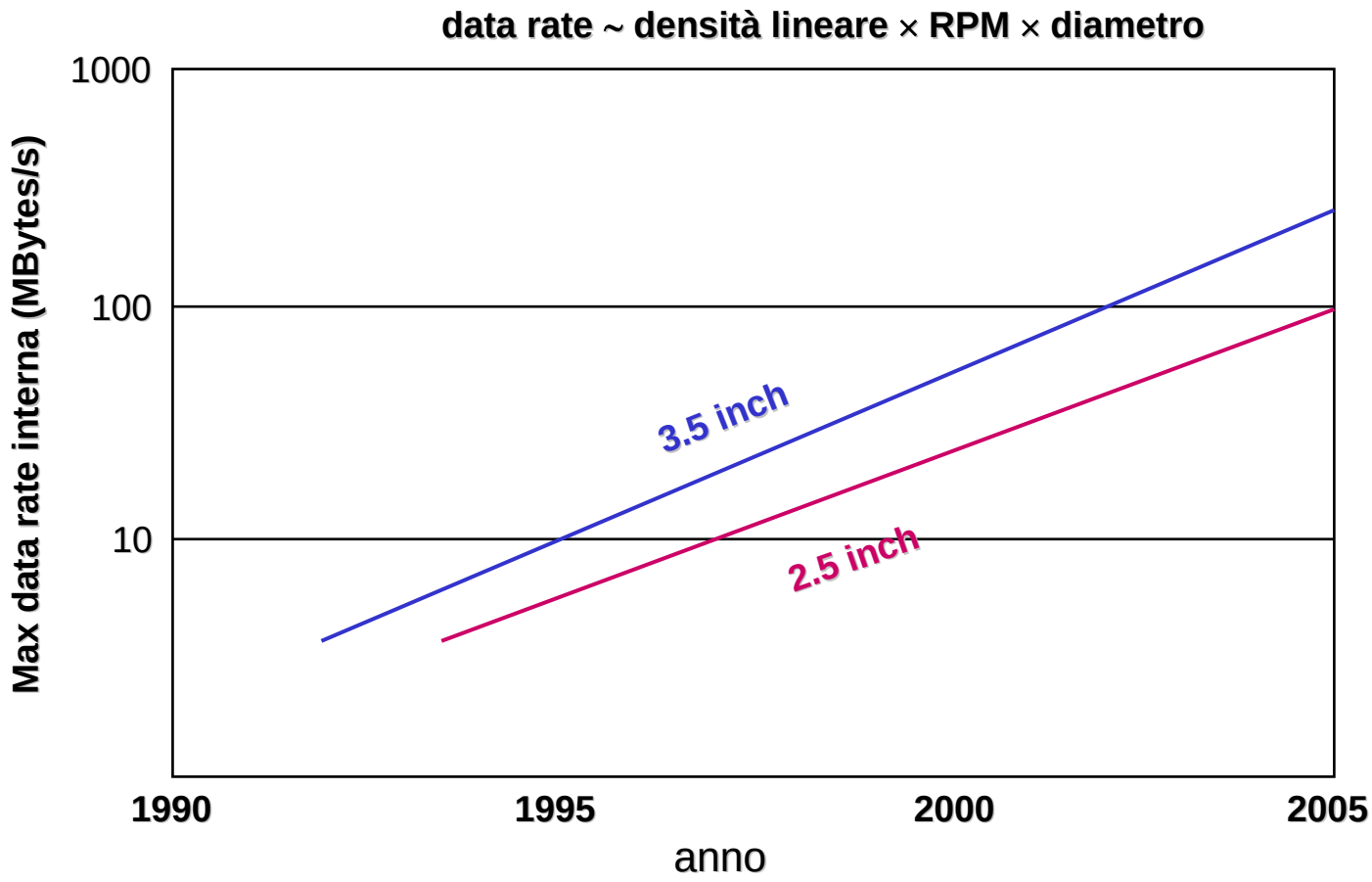
- legge empirica che all'origine riguardava l'andamento della densità dei transistori per chip (che raddoppia ogni 18 mesi)
- concerne l'incremento della *capacità operativa* (che si moltiplica approssimativamente per 100 ogni 10 anni)



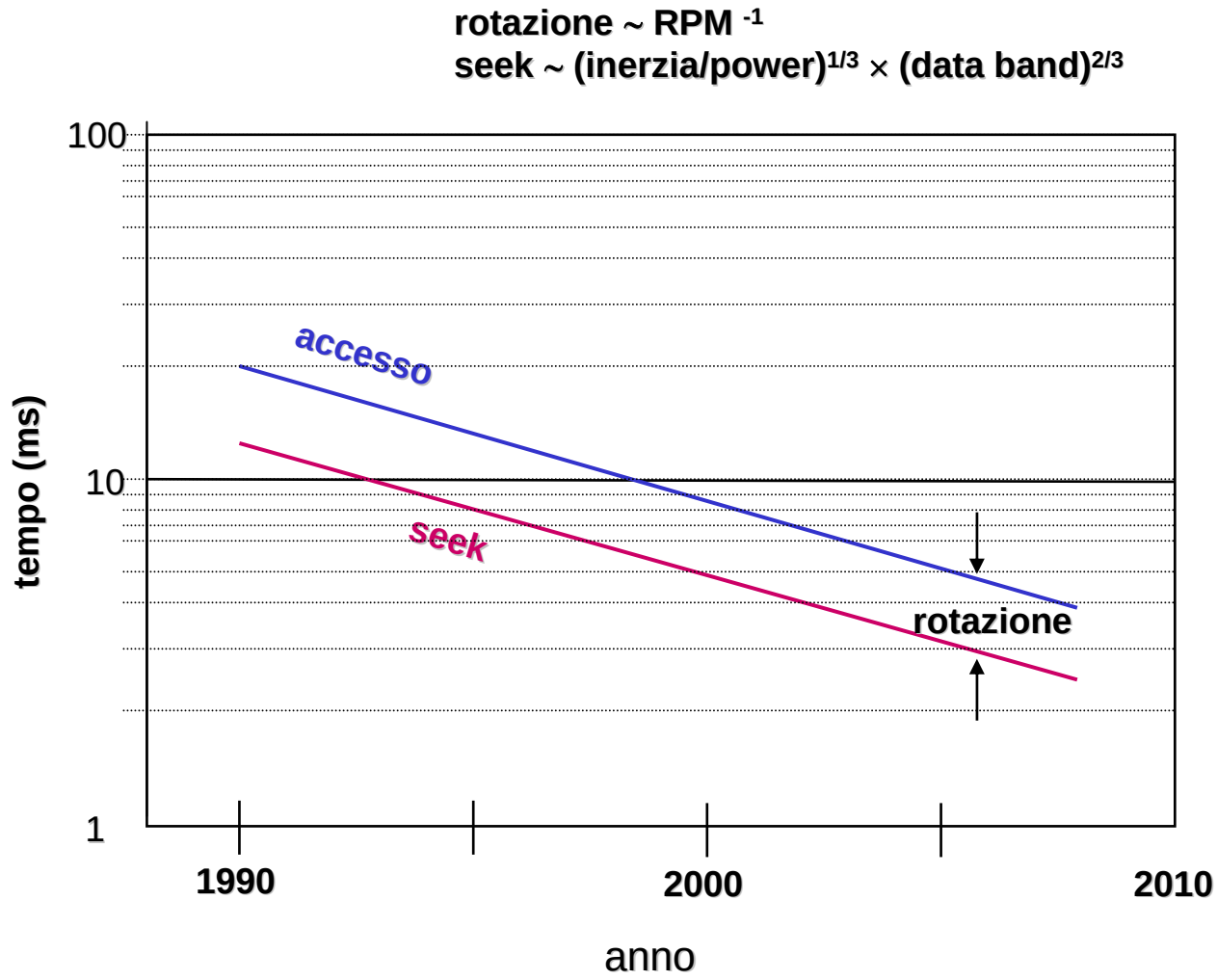
esempio: evoluzione dischi (densità superficiale)



esempio: evoluzione dischi (data rate)

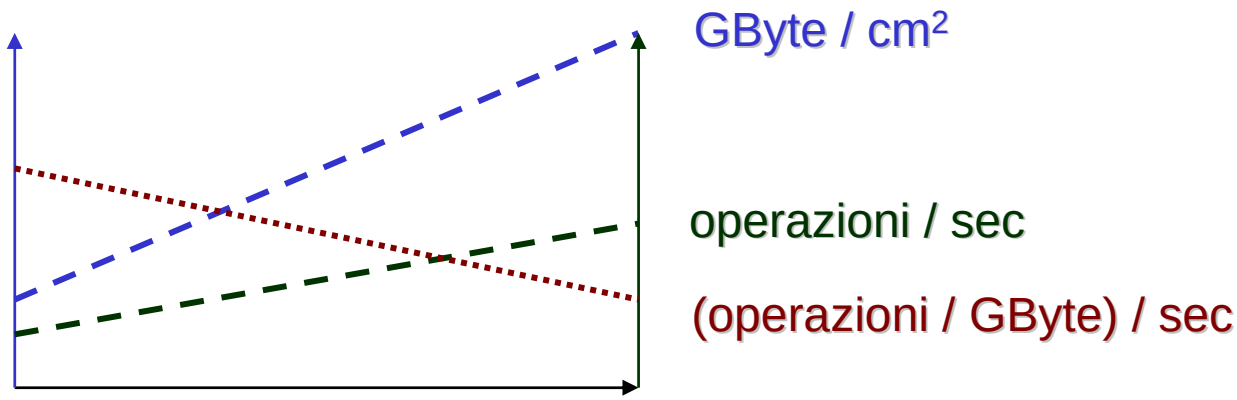


esempio: evoluzione dischi (tempi di accesso)



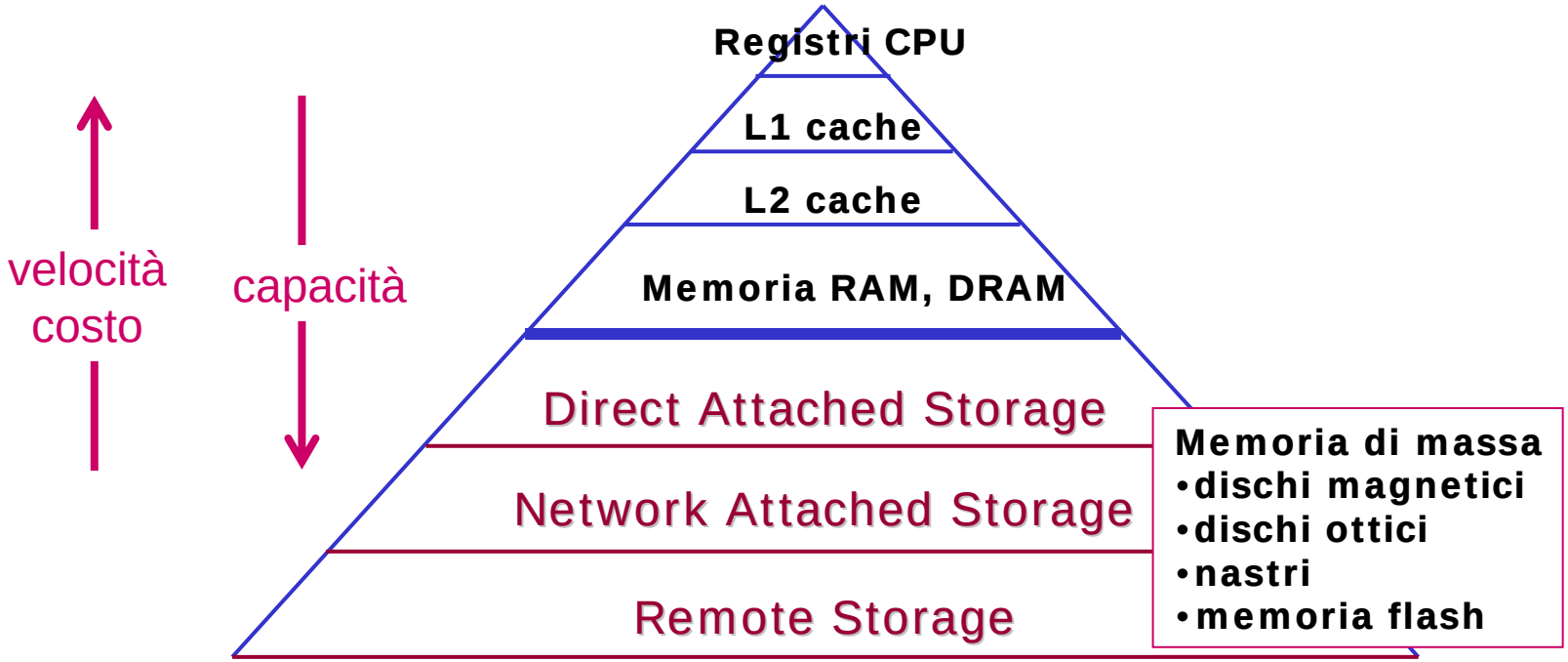
osservazione

- la crescita con tassi diversi di alcune grandezze può dare luogo ad alcuni problemi:
  - in particolare la capacità di memoria cresce più rapidamente dei tempi di accesso perciò la densità degli accessi è diminuita nel tempo
  - pericolo di non completo utilizzo (inedia o "*starvation*") dei dispositivi (processori e dischi)



- la capacità dei dischi è cresciuta dal 1956 al 2005 di 5×10^7 volte,
- entro 4 anni si pensa di raggiungere 500 Gbit per inch^2
- con metodi olografici 1 Tbit può essere contenuto in un volume di 1 cm^3 (*fonte: Scientific American aug. 05*)

livelli gerarchici di memoria



Memorie di massa: Tecnologie

Tecnologie

- Magnetica
- Ottica

Diversi obiettivi:

- Tempo di accesso (msec, nanosec)
- Velocità di trasferimento (Mbyte/sec)
- Capacità (Mbyte/Gbyte)

esercizio 1: calcolo del tempo medio di accesso

Esempio di calcolo del tempo medio di accesso al dato

	layer	*tempo* ***t***	*miss rate* ***m***	*prob.* ***p***	*p x t*
1	*Reg*	1,00E+00	1,00E-01	1,00E+00	1,00E+00
2	*L1*	1,00E+00	5,00E-02	1,00E-01	1,00E-01
3	*L2*	8,00E+00	2,00E-02	5,00E-03	4,00E-02
4	*Main mem.*	1,00E+02	1,00E-01	1,00E-04	1,00E-02
5	*Local disk*	1,00E+07	2,00E-02	1,00E-05	1,00E+02
6	*Net server*	5,00E+07	2,00E-02	2,00E-07	1,00E+01
7	*Remote server*	4,00E+08	0,00E+00	4,00E-09	1,60E+00
				tot	**112,75**

$p(i) = p(i-1) \times m(i-1)$ *probabilità di arrivare nella ricerca del dato al livello (i)*

$p(i) \times t(i)$ *tempo di accesso al livello (i)*

tempo totale $= \Sigma\, p(i) \times t(i)$

osservazione sulle probabilità e loro calcolo

- gli *eventi* E(i) a cui competono le probabilità p(i) sono le visite ai diversi livelli della gerarchia di memoria, *non sono esclusivi* ma:
  - $E(k) \subseteq E(j)$; $p(k) \leq p(j)$ per k > j - la visita al livello (k) richiede di essere prima passati da (j)
  - $\Sigma\ p(i) > 1$
- passiamo a considerare gli *eventi incompatibili* E*(i):
  - E*(i): il dato cercato si trova al livello (i)
  - $p^*(i) = (1- m(i)) \times p(i)$ - probabilità che la ricerca si arresti al livello (i)
  - $\Sigma\ p^*(i) = 1$
  - t*(i): tempo per ottenere il dato residente al livello (i) = somma dei tempi di accesso ai livelli da 1 a (i)
- la tabella della pagina che precede si modifica in quella che segue
- (ovviamente il tempo medio non cambia)

esercizio 2

Esempio di calcolo del tempo medio di accesso al dato (variazione)

	layer	*tempo* ***t***	*miss rate* ***m***	***p****	***t****	***p** x t**
1	*Reg*	1,00E+00	1,00E-01	9,00E-01	1,00E+00	9,00E-01
2	*L1*	1,00E+00	5,00E-02	9,50E-02	2,00E+00	1,90E-01
3	*L2*	8,00E+00	2,00E-02	4,90E-03	1,00E+01	4,90E-02
4	*Main mem.*	1,00E+02	1,00E-01	9,00E-05	1,10E+02	9,90E-03
5	*Local disk*	1,00E+07	2,00E-02	9,80E-06	1,00E+07	9,80E+01
6	*Net server*	5,00E+07	2,00E-02	1,96E-07	6,00E+07	1,18E+01
7	*Remote server*	4,00E+08	0,00E+00	4,00E-09	4,60E+08	1,84E+00
				Σ ***p* = 1***	***tot***	**112,75**

$p^*(i) = p(i) \times (1-m(i))$ *probabilità di trovare il dato al livello (i)*

$t^*(i)$ *tempo per raggiungere il livello (i)*

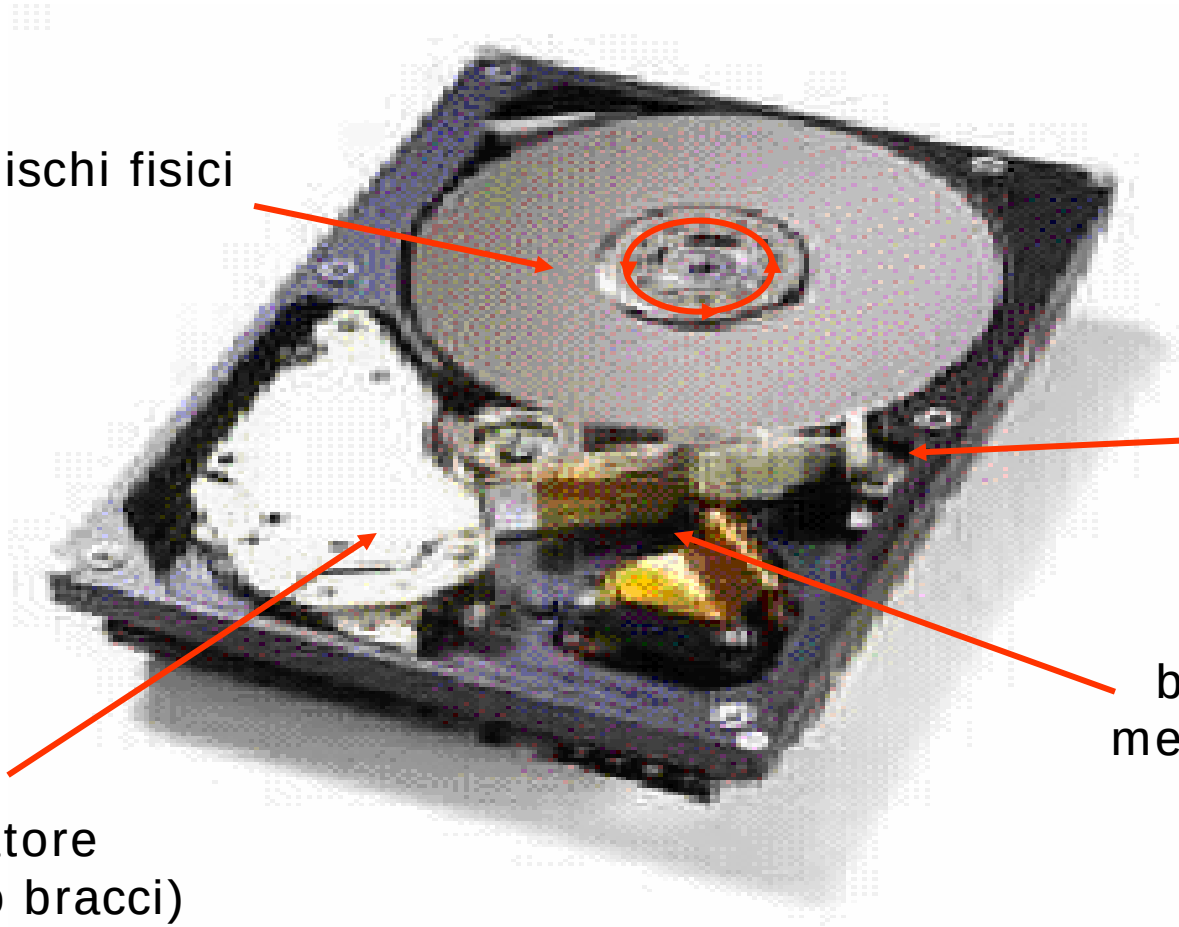
tempo totale $= \Sigma\ p^*(i) \times t^*(i)$

esercizio 3

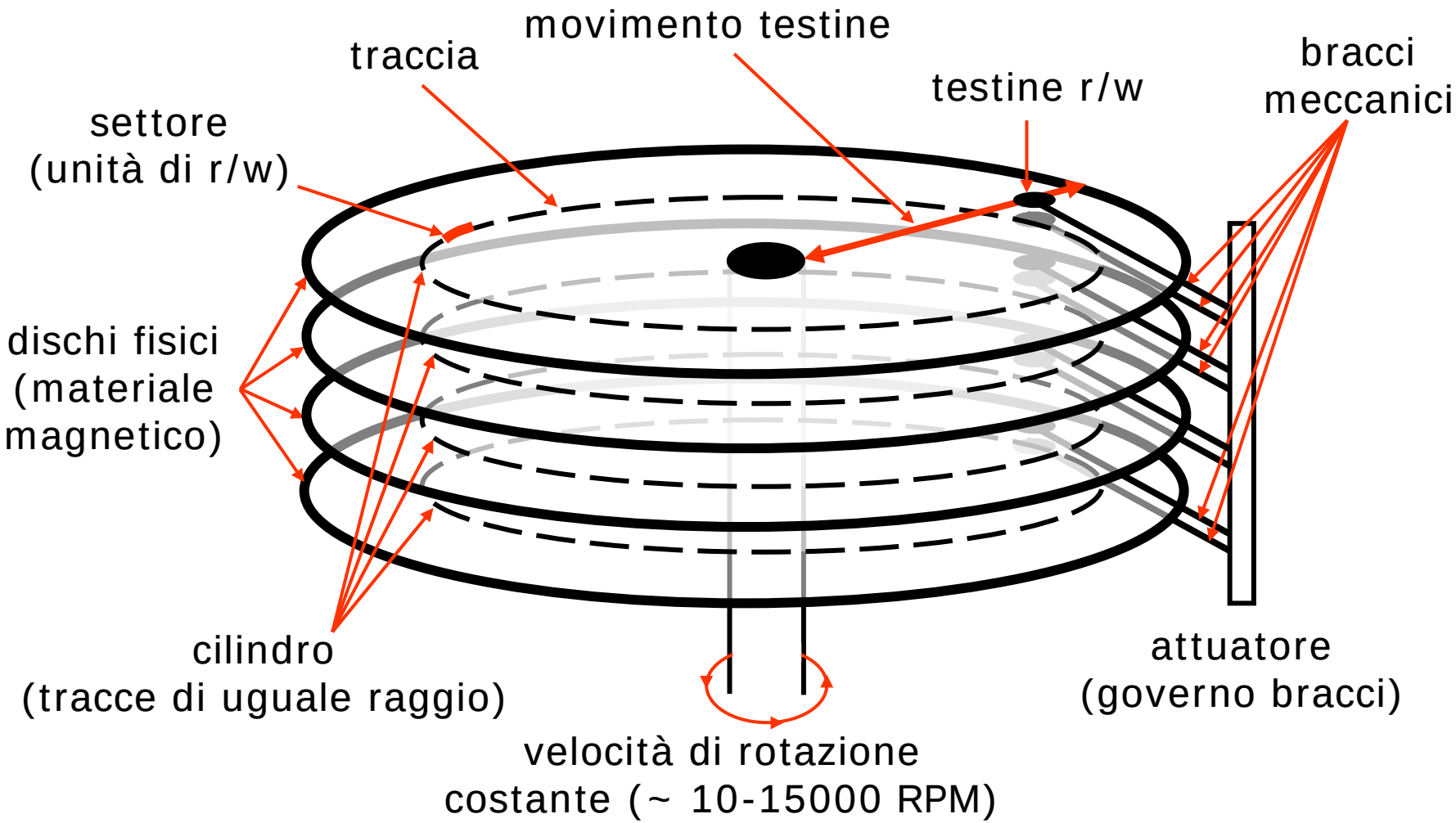
Calcolo del tempo medio di accesso con alcuni valori delle miss rate modificati

	layer	*tempo* ***t***	*miss rate* ***m***	*prob.* ***p***	*p x t*
1	*Reg*	1,00E+00	1,00E-01	1,00E+00	1,00E+00
2	*L1*	1,00E+00	**5,50E-02**	1,00E-01	1,00E-01
3	*L2*	8,00E+00	**2,20E-02**	5,50E-03	4,40E-02
4	*Main mem.*	1,00E+02	**1,10E-01**	1,21E-04	1,21E-02
5	*Local disk*	1,00E+07	**2,20E-02**	1,33E-05	1,33E+02
6	*Net server*	5,00E+07	**2,20E-02**	2,93E-07	1,46E+01
7	*Remote server*	4,00E+08	0,00E+00	6,44E-09	2,58E+00
				tot	**151,47**

unità disco


dischi fisici
testine r/w
bracci
meccanici
attuatore
(governo bracci)

dischi magnetici



Dischi magnetici

È composto da un insieme di piatti di alluminio con rivestimento magnetizzabile

Diametro: 3-50 cm (in genere 9cm=3.5inch)

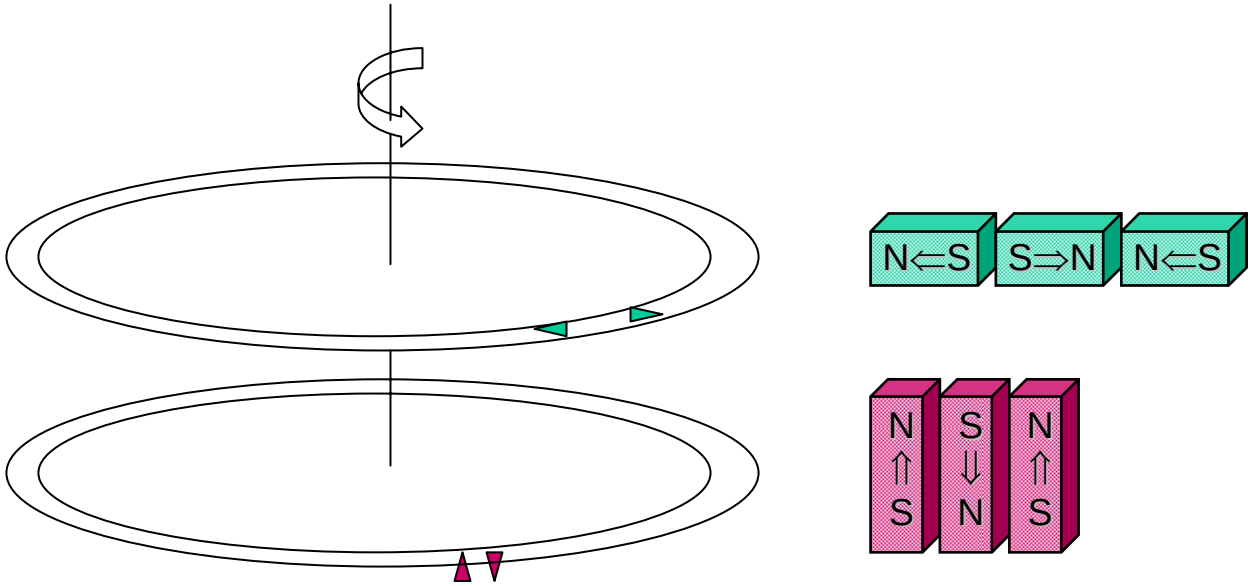
Principi fisici:

- I materiali ferro-magnetici hanno memoria

- Scrittura: la corrente crea un campo magnetico (modifica magnetizzazione)

Lettura: il campo magnetico induce una corrente in alternativa, si misura la magneto-resistenza il cui valore dipende dal campo elettrico

evoluzione della tecnologia

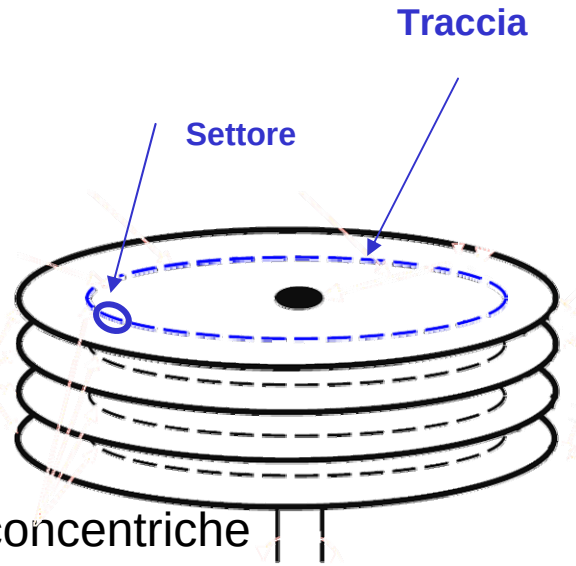


- magnetizzazione **longitudinale** (orizzontale, standard)
- magnetizzazione **perpendicolare** (verticale, si raggiunge una densità × 10)
- ogni ***bit*** contiene da 50 a 100 grani di materiale ferromagnetico magnetizzato
- heat-assisted recording: (ancora sperimentale), permette bit ancora più piccoli di nuovi materiali, magnetizzati con l'intervento laser che riduce momentaneamente la coercività
- *fonte: Scientific American, sept. 2006*

Formattazione in tracce e settori

Piatto:

- Velocità di rotazione: 5400 – 15000 giri/minuto
- Composto da due distinte facce



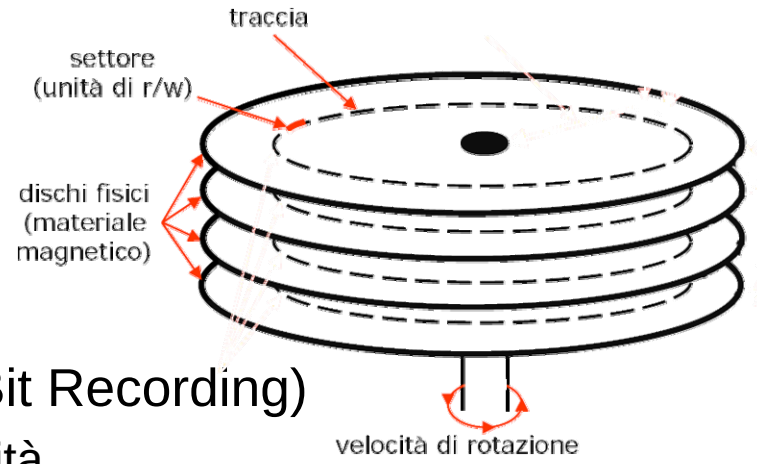
Formattazione:

- Aggiunta di informazioni di controllo
  - Preambolo
  - ECC (Error Correction Code)
  - Spazi vuoti
- Suddivisione della superficie del disco
  - Tracce (track): partizionata in settori
    - ogni faccia contiene 10000-50000 tracce concentriche
  - Settori (sector): contiene un blocco di dati
    - (100-500) tipicamente di 512 byte
    - contengono *numero settore - gap- informazioni con sequenza di correzione errori - gap - numero prossimo settore* e così via

Formattazione in tracce e settori

Numero di settori per traccia

- Costante
  - Semplice da gestire
  - Sottosfruttamento della capacità ideale



- Diverso nelle diverse zone (Zone Bit Recording)
  - Sfruttamento intensivo della capacità
  - Velocità di trasferimento più veloce ai bordi
    - spostandosi verso l'esterno aumenta la capacità delle tracce e quindi la velocità di trasferimento

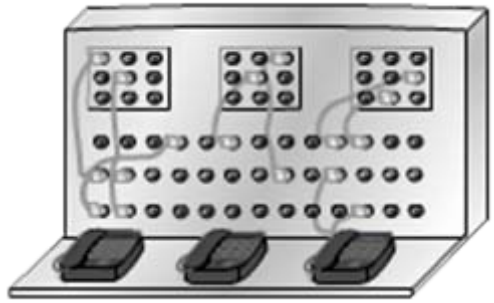
Testine

- le *testine* di lettura/scrittura
  - si muovono insieme e si trovano contemporaneamente sulla stessa traccia per ogni superficie
    - Una per ogni faccia
    - Sollevata (non a contatto) dalla superficie
  - Attuatore:
    - Ogni testina è sorretta da un braccio meccanico
    - Manovra i bracci meccanici che si muovono in modo solidale
  - *cilindro*: insieme delle tracce di tutte le facce accessibili dalle testine posizionate in un determinato punto
  - *seek*: operazione meccanica di posizionamento della testina, la sua durata dipende dalle tracce attraversate, mediamente 3-14 ms
- *trasferimento*: la sua durata è il tempo necessario per leggere/scrivere un blocco
  - dipende da: dimensione del settore, velocità di rotazione e densità dei dati sul supporto, 30-80 MB/sec.
  - buona parte delle unità di controllo dei dischi hanno una cache (velocità di trasferimento fino a 320 MB/sec)
    - ovviamente se il trasferimento avviene dalla cache non c'è né seek né latenza

Controllore del disco

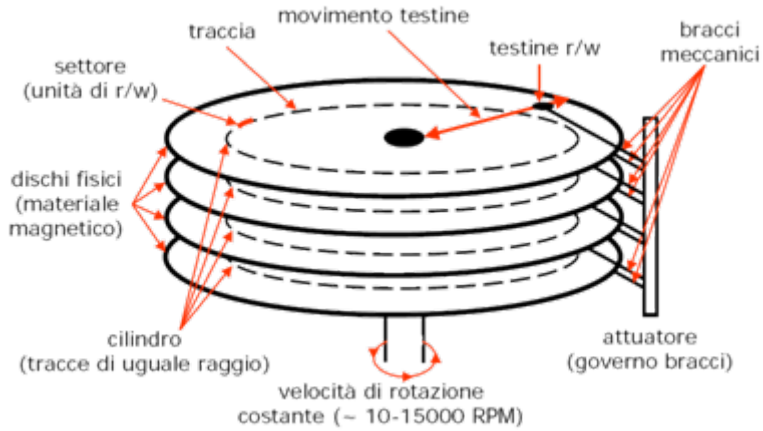
Circuito di controllo:

- Può contenere un processore



Funzionalità:

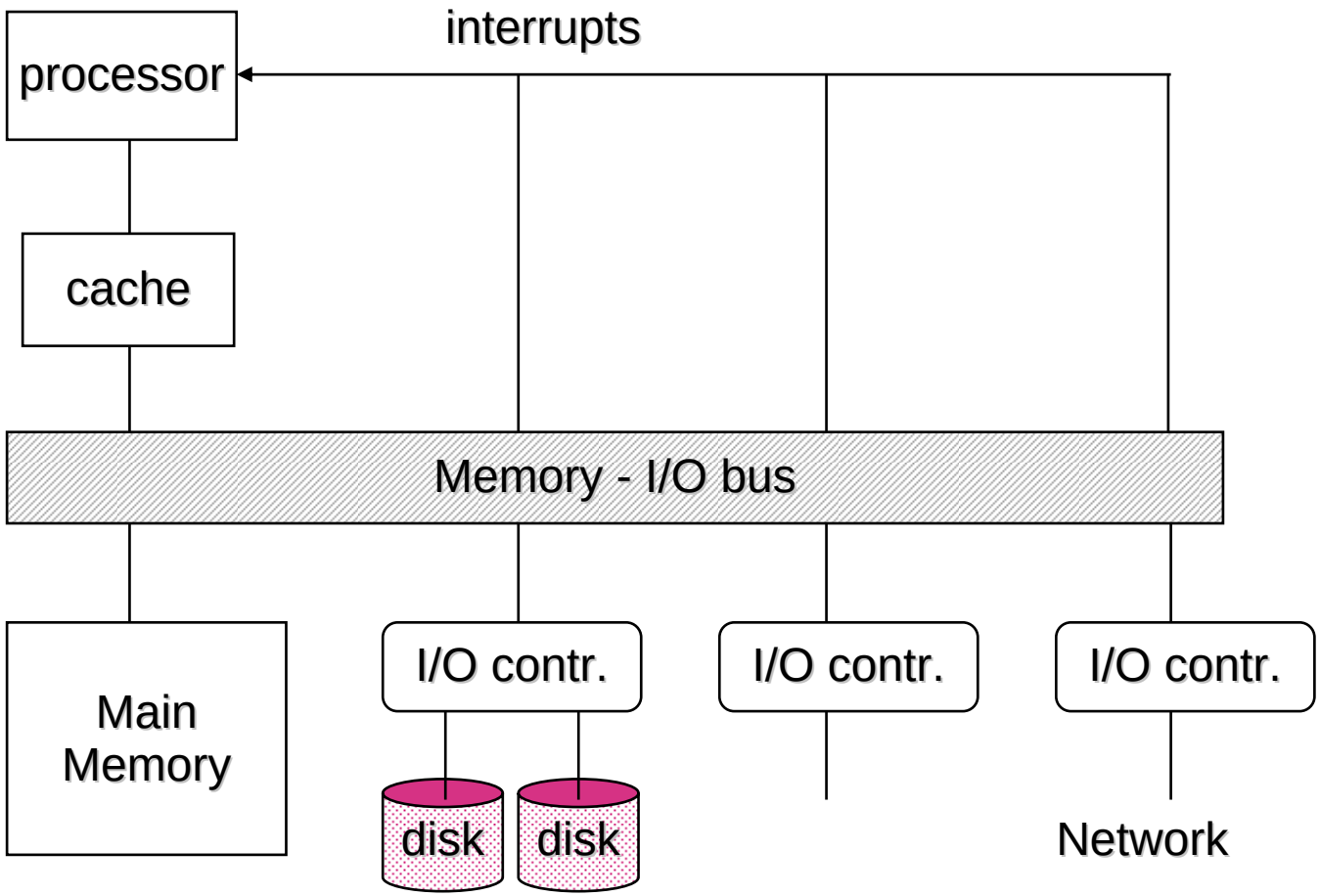
- Movimento della testina
- Correzione degli errori
- Buffering/caching
- Rimappaggio settori danneggiati
- Interfaccia con il bus
  - IDE/ATA
  - SATA
  - SCSI



caratteristiche indicative di dischi (2004)

Characteristics	***Seagate ST373453***	***Seagate ST3200822***	***Seagate ST94811A***
Disk diameter (inches)	2.50	3.50	3.50
Formatted data capacity (GB)	73.4	200	40.0
Cylinders	31310		
Sectors per drive	143,374,744	390,721,968 *(LBA mode)*	78,140,160 *(LBA mode)*
Number of disk surfaces (heads)	8	4	2
Rotation speed (RPM)	15,000	7200	5400
Internal disk cache size (MB)	8	8	8
External interface, bandwidth (MB/sec)	Ultra320 SCSI, 320	Serial ATA, 150	Ultra ATA, 100
Sustained transfer rate (MB/sec)	57-86	32-58	34
Minimum seek (read/write) (ms)	0.2/0.4	1.0/1.2	1.5/2.0
Average seek (read/write) (ms)	3.6/4.0	8.5/9.5	12.0/14.0
Mean time to failure (MTTF) hours	1,200,000@25 ^{o}C	600,000@25 ^{o}C	330,000@25 ^{o}C
Warranty (years)	5	3	-
Nonrecoverable read error per bit read	< 1 per 10^{15}	< 1 per 10^{14}	< 1 per 10^{14}
Price in 2004 ($/GB)	$5	$0.5	$2.5

sistemi di I/O



bus

- *bus*: link di comunicazione condiviso che usa un insieme di *fili* per connettere sottosistemi multipli
  - *vantaggi*:
    - basso costo
    - versatilità
  - unico schema di connessione
    - si possono aggiungere nuovi dispositivi
    - le periferiche che usano lo stesso tipo di bus possono essere spostate fra sistemi diversi
  - *svantaggi*:
    - collo di bottiglia delle comunicazioni, la larghezza di banda limita il massimo throughput I/O
    - lunghezza del bus
    - numero di dispositivi
  - *prospettive*:
    - interconnessioni seriali ad alta velocità con switch

bus (2)

- tipi di bus
  - *processore - memoria* (corti e ad alta velocità)
  - di *I/O* (lunghi con ampio range di banda)
  - *backplane* (permette la coesistenza di memoria e I/O)
  - sono organizzati in modo gerarchico
- comunicazione
  - *sincrona* (generalmente usata dal bus di memoria); i dispositivi devono avere lo stesso ritmo di clock
  - *asincrona* (usata dai bus di I/O); il coordinamento della trasmissione è gestito da un protocollo di *handshaking*; richiede linee addizionali per i segnali di controllo
- trasferimento
  - *parallelo*
  - *seriale*

bus (3)

- ISA
- Micro channel
- EISA
- VESA
- PCI
- AGP
- USB
- IDE
- Serial ATA
- PCMCIA
- SCSI
- Firewire
- Bluetooth

Key characteristics of two dominant I/O bus standards		
Characteristic	***Firewire (1394)***	***USB 2.0***
Bus type	I/O	I/O
Basic data bus width (signals)	4	2
Clocking	asynchronous	asynchronous
Theoretical peak bandwidth	50 MB/sec (Firewire 400) 100 MB/sec (Firewire 800)	0.2 MB/sec (low speed) 1.5 MB/sec (full speed) 60 MB/sec (high speed)
Hot plugable	yes	yes
Maximum number of devices	63	127
Maximum bus length (copper wire)	4.5 meters	5 meters
Standard name	IEEE 1394, 1394b	USE Implementors Forum

Interfaccia ATA/IDE

ATA (AT Attachment)

IDE (Integrated Drive Electronics)

- EIDE (Extended IDE)

→ **PATA (Parallel ATA)**

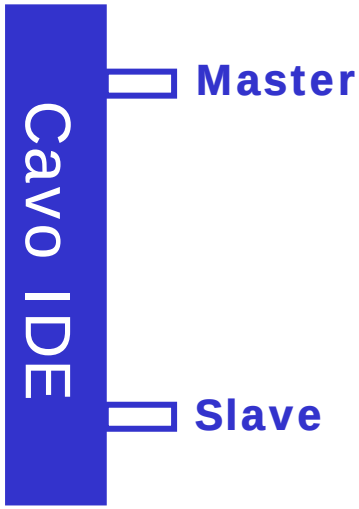
ATA1 – ATA4

- 16 bit
- 40 fili (40 pin)
- 8,3 -33,3 Mbps

ATA5 – ATA6

- 32 bit
- 66 – 133 Mbps
- 80 fili (ma 40 pin)
  - 40 originali
  - 40 x messa a terra

Ultra DMA
(Direct Memory Access)
due cicli di trasferimento per ciclo di clock



Limitazioni:

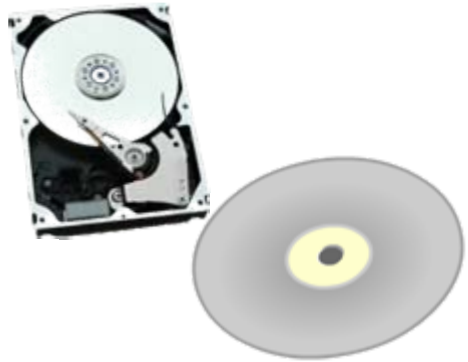
- Transfer rate
- Lunghezza cavo (45 cm) e larghezza

Interfaccia SCSI

SCSI (Small Computer System Interface)

Collega dispositivi eterogenei

- Hard disk
- Unità nastro
- Scanner
- Lettori/masterizzatori CD/DVD
- Stampanti
- …



Adapter SCSI

Controller SCSI

- Incorporato nelle periferiche

Interfaccia SATA

SATA (Serial ATA)

- Evoluzione dell'ATA

Vantaggi:

- Velocità
- Gestione dei cavi
- Hot Swap

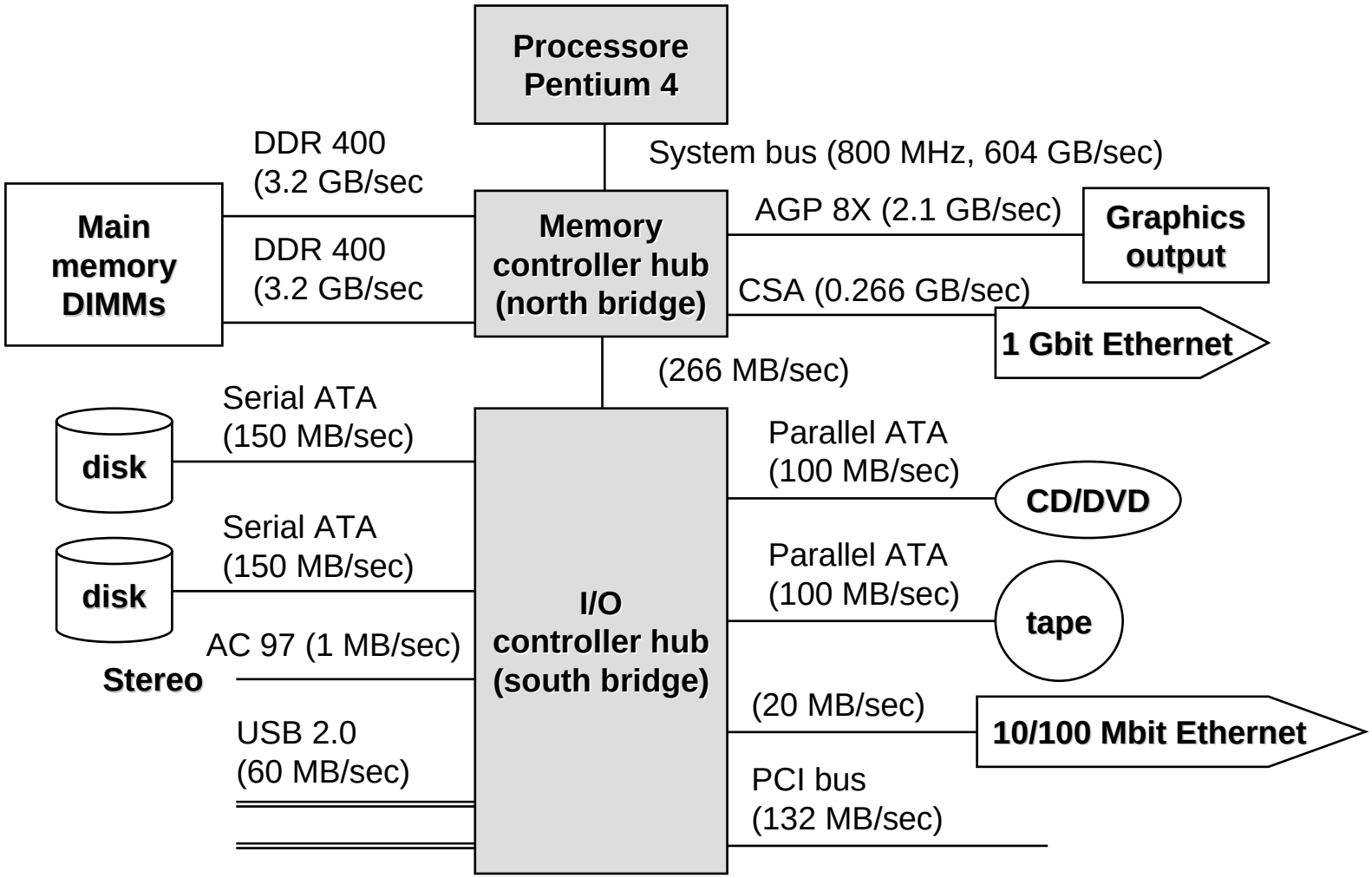
Trasferimento seriale

- No interferenze della connessione parallela
- Facile trasportare un bit alla volta

Caratteristiche:

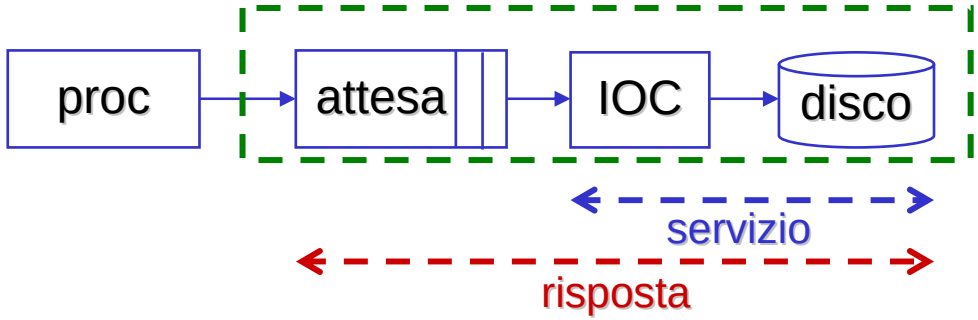
- Cavi a 7 contatti
- Connessione punto a punto
  - Un cavo per ogni dispositivo collegato
- SATA-150
  - Clock 1,50 Ghz
- SATA-300
- SATA-600

organizzazione I/O



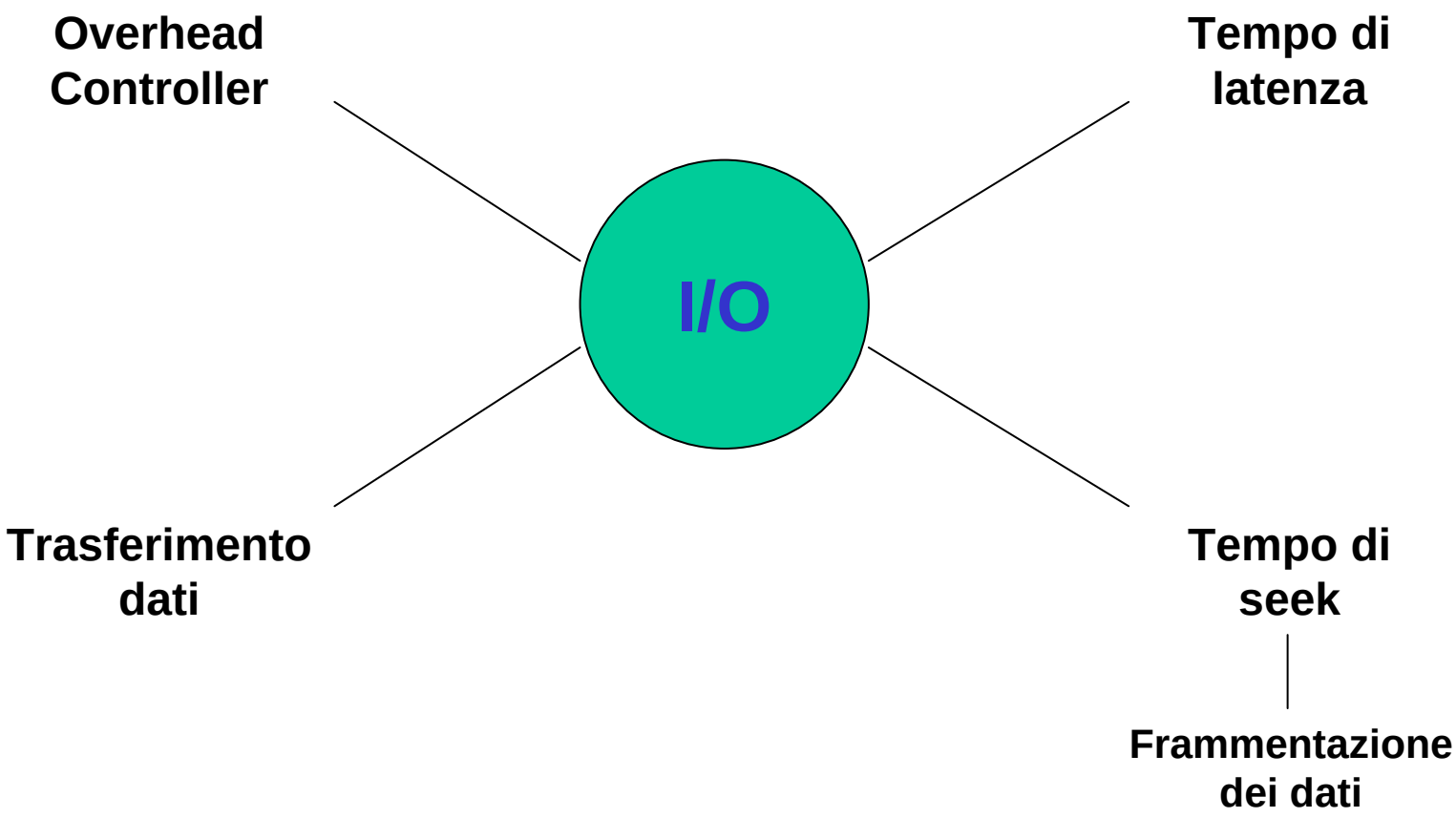
prestazioni dischi

metriche di prestazione di un disco

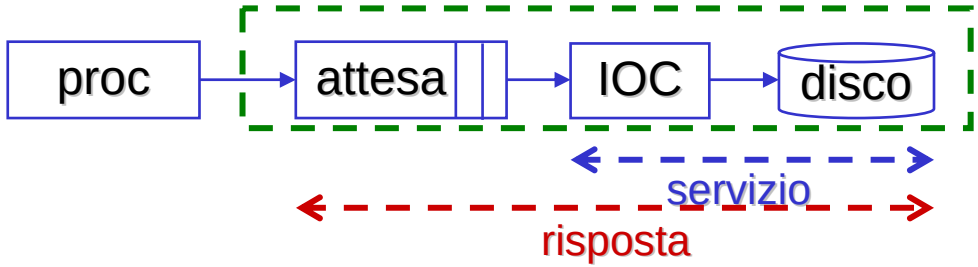


- tempo di servizio s_{disk}: *seek time+rotational latency+data transfer time+overhead controller*
  - tempo di seek: posizionamento testine (≈ ms)
  - tempo di latenza: tempo richiesto affinché il settore passi sotto le testine (≈ ms, ½ giro)
  - tempo di trasferimento dei dati : dipende da velocità di rotazione, densità di registrazione, distanza della testina dal centro del disco (≈ MB/sec)
  - *overhead controller:* gestione trasferimento dati da buffer locale e invio interrupt

Operazione di I/O: tempo di servizio



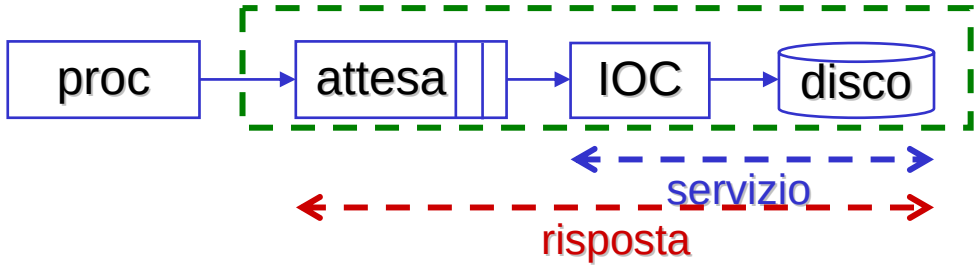
metriche di prestazione di un disco (2)



- il *tempo di risposta r* comprende anche il tempo di attesa causato dalla contesa con altre operazioni concorrenti
- dipende da:
  - numero di utenti trovati in attesa (lunghezza della coda)
  - utilizzo del *servente*
  - tempo medio di servizio
  - variabilità del tempo di servizio (forma della sua distribuzione)
  - tasso di arrivi e sua distribuzione
  - *a parità delle altre condizioni una maggiore variabilità determina un tempo di risposta mediamente più grande*

metriche di prestazione di un disco (3)

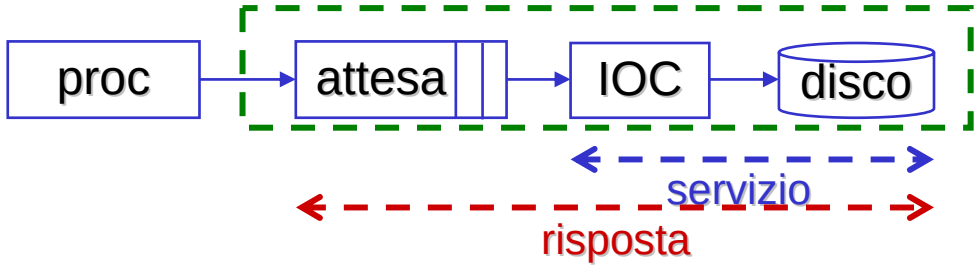
Calcolo delle statistiche
risultanti da composizione
o somma di diversi fenomeni



- *M1: momento del primo ordine; M2: momento del secondo ordine*
- *il tempo di servizio Ti* varia da richiesta a richiesta:
  - *Tempo medio* ti = M1 =(f1×T1+f2×T2+...+fn×Tn)

  (f1+f2+...fn=1)
  - *Varianza* Vi =(f1×$T1^2$+f2×$T2^2$+...+fn×Tn^2) - $M1^2$ = M2 - $M1^2$
  - C^2 = quadrato del coefficiente di variazione
  - C^2 = Varianza / $M1^2$
- *un tempo medio di servizio T* è la somma di tempi medi indipendenti:
  - T = t1+t2+…tm (seek+rotazione+trasferimento+overhead)
  - V = V1+V2+...Vm

metriche di prestazione di un disco (4)



- il *coefficiente di variazione C* = std.dev/M1 è un indice normalizzato della variabilità della distribuzione
- C *cresce all'aumentare della dispersione*
- C = 1 (esponenziale)
  - 63% dei valori < M1; 90% dei valori < 2.3 M1
- C < 1 (ipoesponenziale); C > 1 (iperesponenziale)
- $C^2 = 0.5$
  - 57% dei valori < M1; 90% dei valori < 2 M1
- $C^2 = 2$
  - 69% dei valori < M1; 90% dei valori < 2.8 M1

esercizio 4: tempo medio di servizio di una op. di I/O

- lettura/scrittura di un settore di 512 Byte = 0.5 KB,
  - velocità di rotazione: 10000 RPM (rotazioni per minuto)
  - velocità di trasferimento dati: 50 MB/sec
  - seek medio: 6ms
  - overhead controller: 0.2ms
  - latenza: (60s/min)x1000/(2x10000giri/min)=3.0ms (tempo per compiere mezza rotazione
  - trasferimento: (0.5KB)x1000 / (50x1024KB/s) = 0.01ms
- tempo totale di servizio di I/O = 6ms+ 3ms+ 0.01ms+ 0.2ms = 9.21ms

seek **latenza** **trasferimento** **controller**

esercizio 5: effetti della località degli accessi

- località: effetto come seek nulli
- riprendendo i dati dell'esercizio precedente:
  - località dei dati: seek ha luogo solo nel 25% delle operazioni

$(6.0 \times 0.25) + (0.5 \times 60 \times 10^{3} / 10000) + (0.5\ KB / 50MB \times 2^{10}) + 0.2 =$

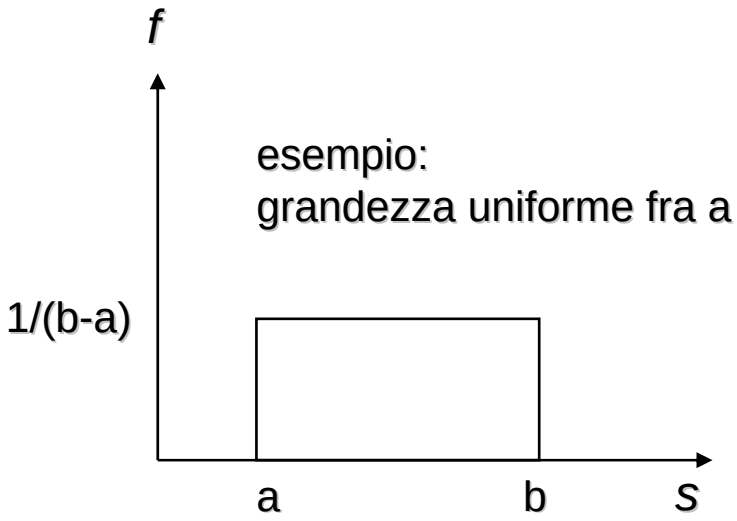
$1.5 + 3.0 + 0.01 + 0.2 = 4.71\ ms$

  - tempo medio = $(0.25 \times 6) + 3 + 0.01 + 0.2 = 4.71\ ms$

seek **latenza** **trasferimento** **controller**

calcolo del tempo di servizio di una op. di I/O

- *tempo medio s = seek + latenza + trasferimento + controller*
- *ipotesi:*
- seek: uniforme fra un valore minimo e un massimo
- latenza di rotazione: uniforme fra 0 e tempo di rotazione completa
- trasferimento: costante
- controller: costante



media M1 = (b+a) / 2

2do Momento M2 = $(b^2 + ab + a^2) / 3$

Varianza V = $(b-a)^2 / 12 = M2 - M1^2$

esercizio 6: Varianza del tempo di servizio di I/O

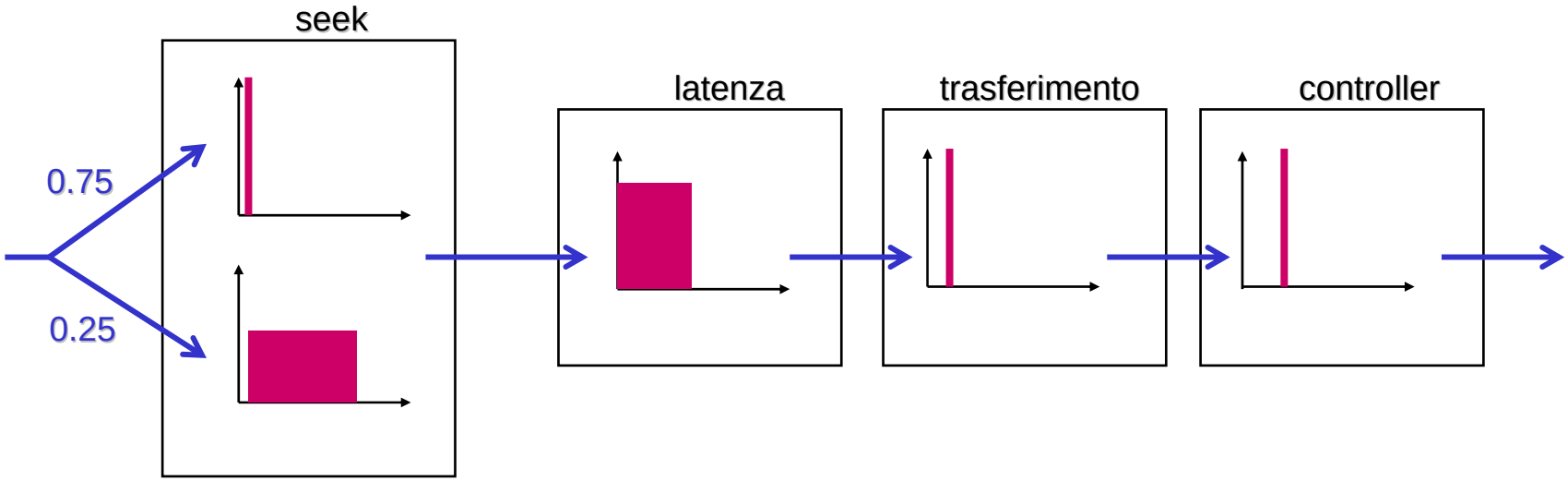
- V(tempo di seek):
  - $M2 = 0.25 \times (11.6^2 + 11.6 \times 0.4 + 0.4^2) / 3 = 0.25 \times 46.45$
  - $V = M2 - M1^2 = 11.61 - (0.25 \times 6)^2 = 9.36$
- V(tempo di latenza):
  - $V = 6^2 / 12 = 3$
- V(tempo di trasferimento):
  - V = 0
- V(tempo di controller):
  - V = 0
- *Varianza totale* = 9.36 + 3 + 0 + 0 = 12.36

i dati provengono dall'esercizio 5

seek **latenza** **trasferimento** **controller**

- *coeff. di variazione* = $\sqrt{(12.36)}/4.71 = 0.747$
- $0.747 < 1 \Rightarrow$ distribuzione ipo-esponenziale

esercizio 6: riassunto grafico



t1 = 0 ; V1 = 0
t2 = 6 ; V2 = 10.45

T(seek) = 1.5
V(seek) = 9.36
C = 2.04

T(latenza) = 3
V(latenza) = 3
C = 0.58

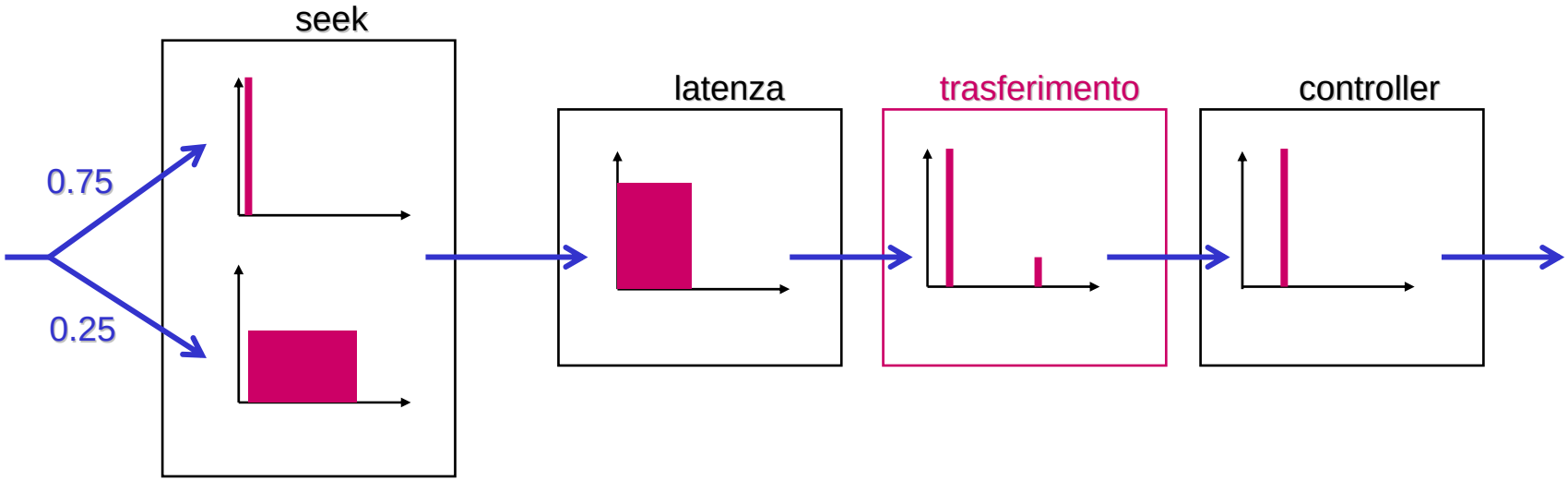
T(trasf.) = 0.01
V(trasf.) = 0
C = 0

T(contr.) = 0.2
V(contr.) = 0
C = 0

il fenomeno complessivo ha un tempo medio $s = \Sigma T = 4.71$

e varianza $= \Sigma V = 12.36$; $C = 0.747$

esercizio 7: variazione sui tempi di trasferimento



t1 = 0 ; V1 = 0
t2 = 6 ; V2 = 10.45

T(seek) = 1.5
V(seek) = 9.36
C = 2.04

T(latenza) = 3
V(latenza) = 3
C = 0.58

T(trasf.) = 0.609
V(trasf.) = 3.23
C = 2.95

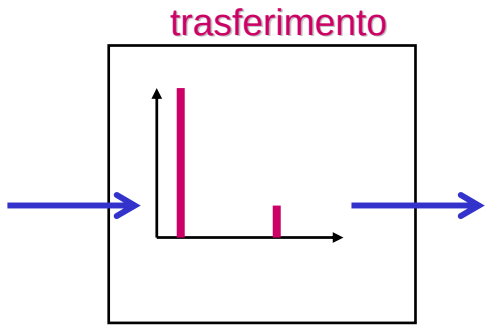
T(contr.) = 0.2
V(contr.) = 0
C = 0

il 90% delle I/O trasferisce 512 byte,
il 10% una intera traccia

il fenomeno complessivo ha un tempo medio s = Σ T = 5.31

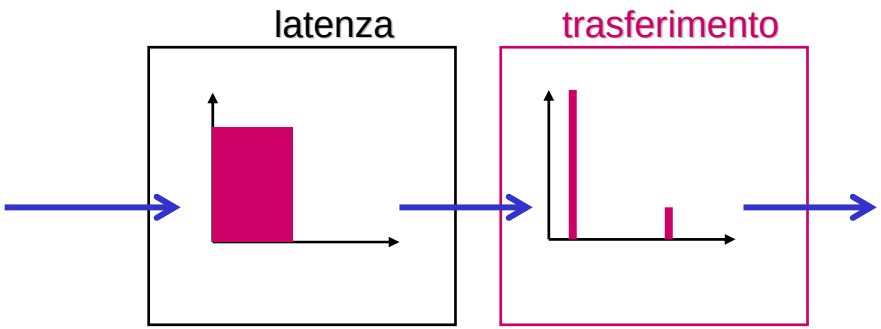
e varianza = Σ V = 15.59; C = 0.744

esercizi 6 e 7: osservazione sulle varianze



Più un fenomeno assume valori diversi
più il coefficiente di variazione tende a crescere

esempio: tempo di trasferimento
t1 = 0.01; **c = 0** nel 90% dei casi
t2 = 6; **c = 0** nel 10% dei casi
t(trasf totale) = 0.609; **c(trasf totale) = 2.95**



se un fenomeno è la somma di più fenomeni diversi,
il coefficiente di variazione totale è minore della media dei singoli coefficienti

esempio: tempo di latenza + trasferimento
t(latenza) = 3; **c(latenza) = 0.58**
t(trasf totale) = 0.609; **c(trasf totale) = 2.95**
t(latenza + trasferimento) = 3.609
c(latenza + trasferimento) = 0.69

$c = \sigma$ / media

esercizio 8: allocazione dati “non ottimale”

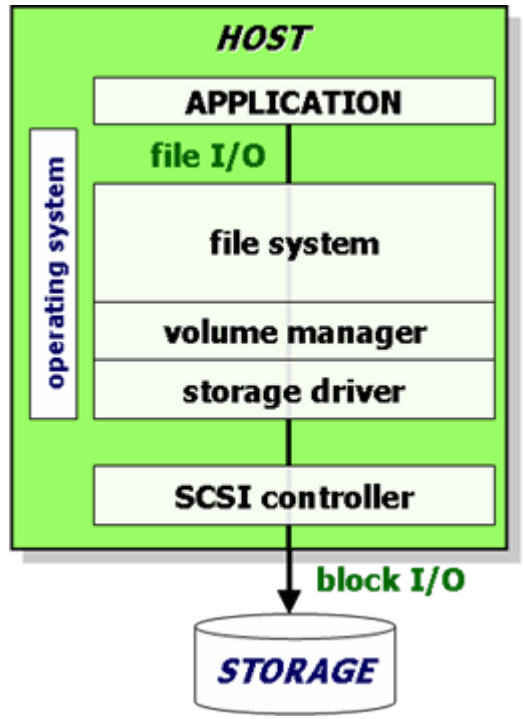
- trasferimento di un file (size = 1MB)
- 1° caso:
  - 1 seek iniziale: 6 ms
  - 1 latenza: 3 ms
  - 1 trasferimento totale 1 MB: $1/50 \times 1000 = 20$ ms
  - tempo totale: 29 ms
- 2° caso, 10 blocchi da 1/10 MB distribuiti “male” sul disco, ciascuno:
  - 1 seek: 6 ms
  - 1 latenza: 3 ms
  - 1 trasferimento parziale: 2 ms
  - tempo totale: $(6 + 3 + 2) \times 10 = 110$ ms

(non si è tenuto conto dei tempi di overhead del controller)

tecniche per migliorare le prestazioni di I/O

tecniche per migliorare le prestazioni I/O

- le prestazioni I/O possono essere misurate sperimentalmente a diversi livelli della gerarchia di storage.
  - per quantificare sperimentalmente gli effetti delle diverse tecniche di ottimizzazione bisogna misurare i tempi da quando la richiesta è consegnata al sistema storage prima che venga potenzialmente spezzata dal *volume manager* in richieste dirette a dischi multipli.



- due metriche importanti sono *response time* e *throughput*.
- il reciproco del *service time* è una stima (ottimistica) del *throughput massimo* (ottenibile con utilizzo = 1)

Ottimizzazione delle prestazioni

- Read Caching
- Prefetching
- Write Buffering

Principio di Località:

Spaziale

temporale

Ottimizzazione delle prestazioni: Read Caching

Read Caching

Prefetching

Write Buffering

I dati più probabili vengono mantenuti in una memoria cache

- Dati: blocchi di 4 Kbyte
- Metodo di alimentazione LRU.

I dischi hanno spesso una loro cache

Read miss ratio

- Frazione di operazioni che richiedono l'accesso al disco
- Dato non in cache
- Indica l'efficienza
- Dimensione della cache: ottimale fino al 4% dello storage totale

Ottimizzazione delle prestazioni: Read Caching

Read Caching

Prefetching

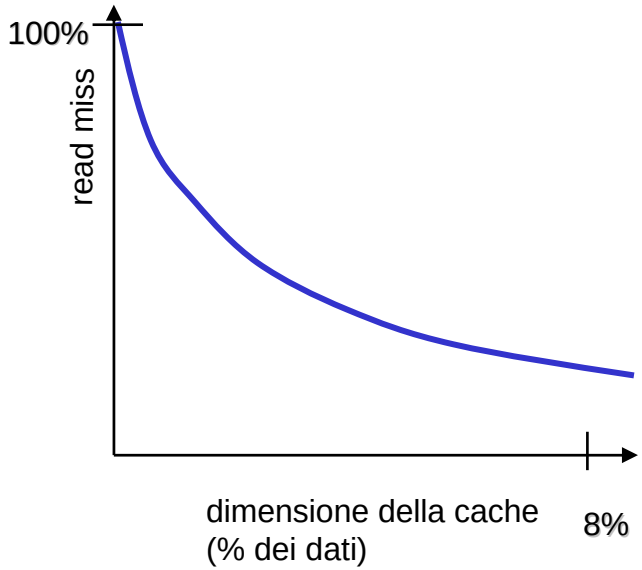
Write Buffering

L'analisi di dati sperimentali mostra una miss ratio che segue la relazione:

$$f(x) = a(x-b)^c \quad (a,b,c \text{ costanti, } c = -1)$$

La relazione funzionale è simile a quella che si trova a livello logico per i buffer dei database

(ma in questo caso il valore di c è circa la metà, in altre parole si trova che il caching a livello fisico è più efficiente di quello ottenuto a livello logico).



Ottimizzazione delle prestazioni: Prefetching

Precarica in memoria i dati che si presume saranno richiesti a breve

- Accuratezza previsione
- Costi di prefetching (risorse consumate)
- Tempestività dell'operazione (operazione completa prima che i dati servano)

Molti accessi al disco sono sequenziali

- Si può realizzare prefetching sequenziale in occasione di una *cache miss*
- Molteplici accessi vengono trasformati in uno singolo

Prefetching: Large fetch unit

Read Caching

Prefetching

Write Buffering

Carica anche i dati precedenti e successivi quelli effettivamente richiesti

Response Time penalty

- Occorre attendere che il trasferimento di tutti i dati sia completato
- Implementazioni “furbe”
  - Si precarica fino al blocco target
  - Si caricano i blocchi rimanenti se non arrivano altre richieste di I/O

Prefetching: Read ahead

Carica i blocchi successivi a quelli richiesti

Si eseguono due operazioni:

- Caricamento blocchi *target*
- Caricamento blocchi successivi

Prefetching: Preemptible read ahead

Usa risorse di sistema altrimenti non sfruttate (idle)

- Si divide una richiesta di prefetching in tante piccole sottorichieste
- Si interrompe il prefetching non appena giunge una nuova richiesta di I/O
- Garantisce buone prestazioni anche al crescere della domanda

Approccio ibrido:

- Si legge la traccia richiesta
- Si prosegue di 32 Kbyte
- Se non ci sono nuove richieste si precarica fino a 128 Kbyte

Prestazioni:

- Server: + 50% rispetto a sistema senza prefetching

Ottimizzazione delle prestazioni: Write Buffering

Mantiene temporaneamente in memoria i blocchi da scrivere sullo storage

Nasconde e differisce il tempo di latenza

Rischio di perdita dei dati

- Buffer su memoria non volatile (NVS)
- Buffer trasferito su disco periodicamente

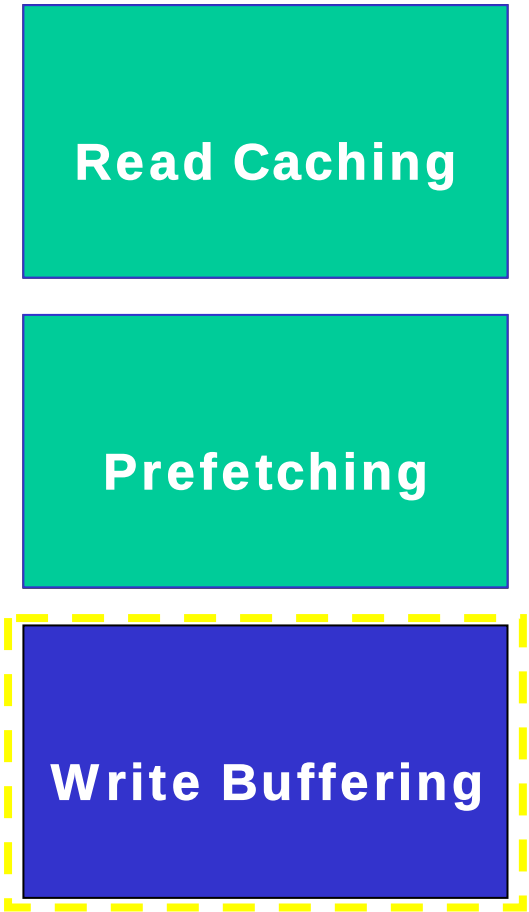
Efficienza

- Write multiple in un'unica operazione

Meno operazioni fisiche

- Una write fisica combina operazioni multiple sulla stessa posizione

Write Buffering



Si effettua il *destage* dei blocchi che non necessiteranno di riscritture

- Si inizia quando il numero di blocchi modificati supera l'*high mark*
- Si termina quando il numero di blocchi modificati è minore del *low mark*

Il blocco da *scaricare* (*destaging*) è selezionato

- Con il metodo *Least-recently-written* (LRW)
- Quando supera un massimo di età
- Traccia con maggior numero di blocchi modificati

Write Buffering

Read Caching

Prefetching

Write Buffering

Equilibrio tra eliminazione delle write e scrittura multipla

- *High mark* = 0.8
- *Low mark* = 0.2

Se *Low mark* << *High mark* il *destage* avviene a lotti

Per minimizzare il tempo di attesa si possono ordinare le write fisiche, in base a:

- Tempo minimo di accesso
- Tempo minimo di posizionamento

Write Buffering

Read Caching

Prefetching

Write Buffering