Sistemi RAID

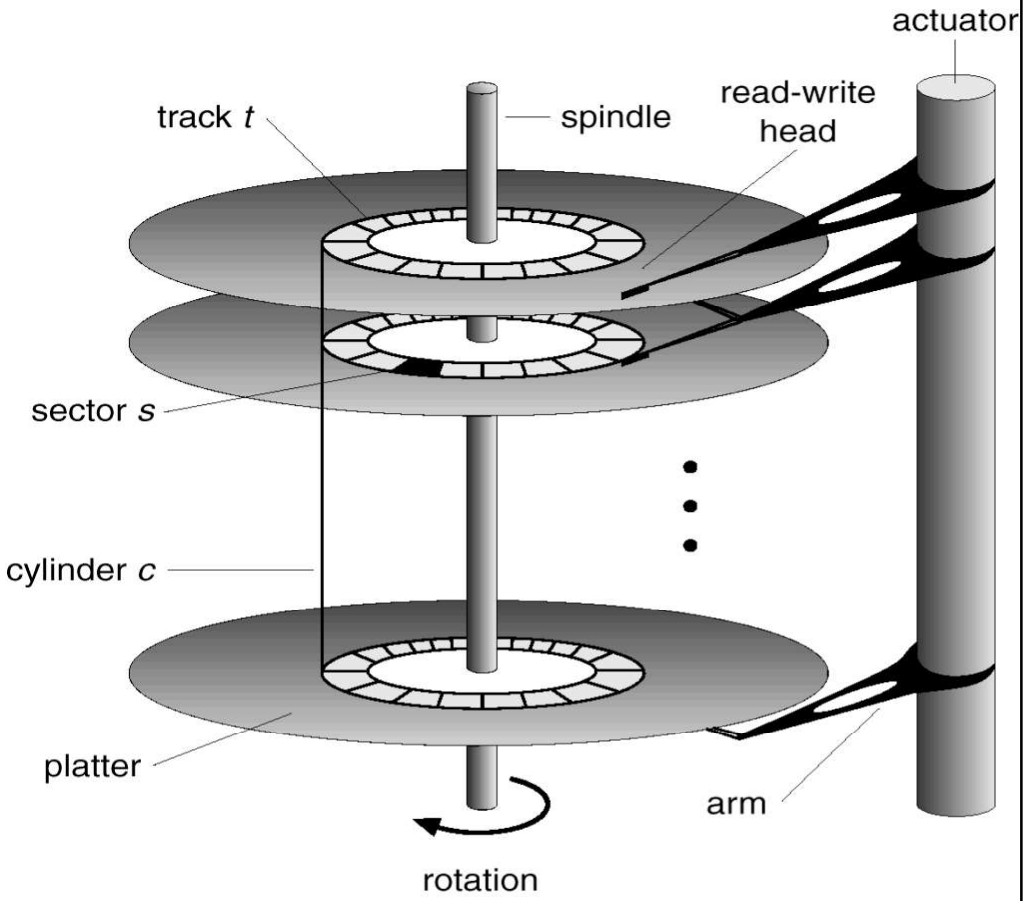
Ripasso

- Ricordiamo: un sistema operativo vede un disco come formato da **blocchi logici**, numerati consecutivamente partire da 0.
- I file e le directory del file system vengono memorizzati in uno o più blocchi logici, secondo lo schema di allocazione adottato: contiguo, concatenato, indicizzato. Per recuperare tutti i dati di un file il SO deve sempre poter risalire al numero di tutti i blocchi logici in cui ha suddiviso il file.
- Tuttavia, fisicamente un hard disk è diviso in piatti o dischi magnetici sovrapposti, ogni piatto è sudddiviso in traccie circolari concentriche, e ogni traccia è suddivisa in settori da 512 byte (o alternativamente, in settori da 256 o 1024 byte: questo valore viene scelto all'atto della formattazione a basso livello del disco, normalmente fatta dal produttore del disco).

Ripasso

- La struttura interna di un hard disk: i dischi sono in continua rotazione (Silberschatz, fig. 12.1)

- Ciascun settore memorizza esattamente un blocco logico (insieme ad alcune informazioni aggiuntive, come il codice di controllo degli errori): il controller di un normale disco (uno SLED) mappa ciascun blocco logico su un ben preciso settore del disco.

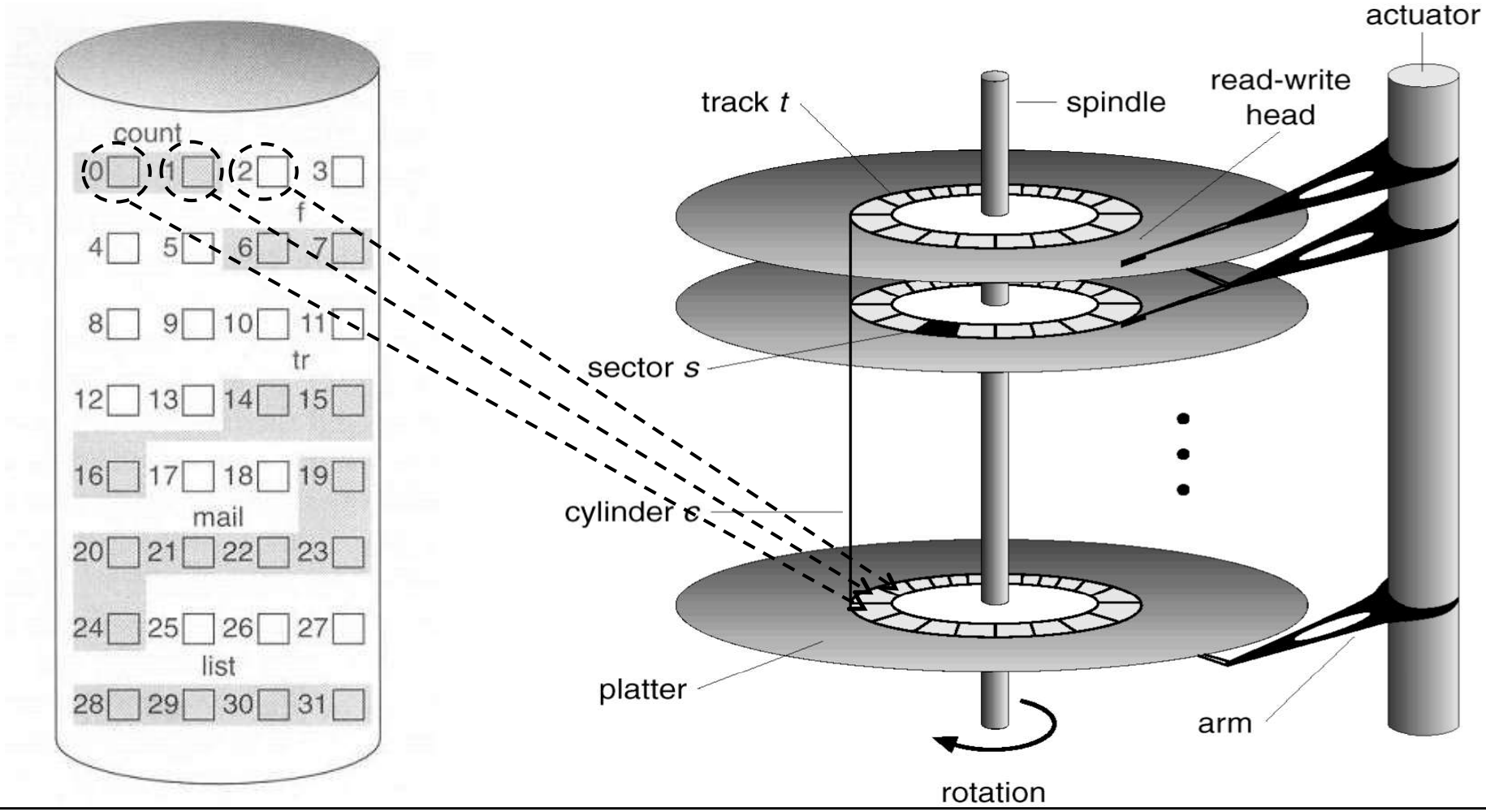


Ripasso

- Normalmente, il blocco logico 0 viene mappato su uno dei settori della traccia più esterna del primo piatto, il blocco logico 1 viene mappato sul settore immediatamente adiacente (percorrendo la traccia nel senso stabilito dal controller), e così via. Quando i settori di una traccia sono finiti si passa alle tracce più interne, e poi si prosegue in modo analogo con i settori dei piatti successivi.
- Quindi, blocchi logici numerati consecutivamente sono memorizzati in settori adiacenti, eccetto quando si passa da una traccia alla successiva, e da un piatto al successivo.
- Quando il SO vuole accedere ad un blocco logico (ad esempio per leggere una porzione di file) presenta il numero del blocco al controller del disco, il quale usa il numero per risalire al settore che contiene il blocco, legge il settore e lo trasferisce in RAM (attraverso il DMA) a partire dalla locazione specificatagli dal SO

Sistema Operativo e SLED

- Quindi, riassumendo, il SO vede un disco virtuale con blocchi numerati consecutivamente che vengono mappati sui vari settori (anch'essi consecutivi) dello SLED dal controller del disco.



Sistemi RAID

- Dei tre elementi fondamentali di un qualsiasi sistema computerizzato: processore, memoria primaria, memoria secondaria, quest'ultimo è di gran lunga il più lento.
- Inoltre, il guasto di un hard disk è potenzialmente il più dannoso: se si guasta, **tutti i dati che contiene** -- e non solo i dati in corso di computazione -- possono venire persi (o per lo meno non essere disponibili per il tempo necessario alla riparazione)
- Un sistema **RAID** è una configurazione della memoria secondaria che permette di aumentare le prestazioni degli hard disk e/o la loro affidabilità.
- Queste "architetture" di memoria secondaria, utili in qualsiasi settore, sono addirittura fondamentali in quei contesti in cui il servizio fornito non può mai venir meno, ad esempio in campo finanziario e bancario.

Sistemi RAID

- Il concetto di dispositivo RAID è stato introdotto nel 1988 da Patterson, Gibson e Katz, e si è velocemente affermato a livello commerciale.
- L'acronimo inizialmente coniato da Patterson stava per **Redundant Array of Inexpensive Disks,** presto però ridefinito dalle varie compagnie produttrici di sistemi RAID in **Redundant Array of Independent Disks**
- In modo un po' simile alla contrapposizione RISC / CISC, fu definita anche la controparte dei sistemi RAID, indicata con l'acronimo **SLED**: **Single Large Expensive Disk**.

Sistemi RAID

- Un sistema RAID è composto da un insieme di hard disk (un *disk array*) ma viene visto dal sistema operativo che lo usa come un normale disco singolo (dovremmo dire: come uno SLED), tuttavia **più veloce** ed **affidabile** di uno SLED.
- In particolare, un sistema RAID è normalmente costituito da uno **SCSI controller** e da un insieme di dischi **SCSI** (Small Computer System Interface).
- la logica interna al RAID organizza l'uso dei vari dischi (secondo diversi criteri che vedremo più avanti) come un unico dispositivo di memorizzazione.
- La possibilità di usare un sistema RAID come se fosse un normale hard disk SCSI fa si che non siano necessari cambiamenti software nel sistema operativo che deve usare il RAID (il che è ovviamente un vantaggio, specie per i system administrators del sistema)

Sistemi RAID

- le idee di fondo di un sistema RAID sono sostanzialmente due:

1. **distribuire l'informazione memorizzata su più dischi**, in modo da parallelizzare una parte delle operazioni di accesso ai dati e guadagnare in prestazioni.
2. **Duplicare l'informazione memorizzata su più dischi**, in modo che in caso di guasto di un disco sia possibile comunque mantenere funzionante il sistema, recuperando in qualche modo l'informazione memorizzata sul disco guasto.

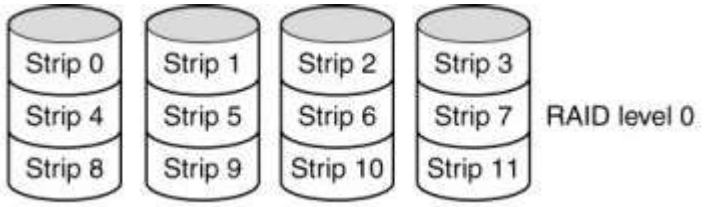
- Differenti schemi sono stati proposti per realizzare i punti 1 e 2, a cui corrispondono diversi livelli di sistemi RAID: dal livello 0 al livello 6.

RAID livello 0 (block striping)

- I sistemi RAID di livello 0 non sono in senso stretto dei sistemi RAID, in quanto non c'è nessuna duplicazione dei dati.
- In un RAID livello 0, il disco virtuale (ossia ciò che viene visto dal sistema operativo: un insieme di blocchi logici numerati consecutivamente) viene mappato dalla logica del RAID sui vari settori dei vari dischi di cui è composto il sistema suddividendo i blocchi logici del disco virtuale in **strips** (strisce) di k blocchi consecutivi ciascuna.
- Quindi, lo strip 0 contiene i blocchi logici da 0 a k - 1, lo strip 1 contiene i blocchi logici da k a $2k$ - 1, e così via.

RAID livello 0 (block striping)

- Il RAID controller suddivide poi gli strip tra i dischi del sistema secondo la formula: *numero-strip* MOD *dischi-nel-sistema*.
- Ad esempio, se sono disponibili 4 dischi, il disco 0 conterrà gli strip 0, 4, 8,... il disco 1 conterrà gli strip 1, 5, 9,..., e così via (Tanenbaum, Fig. 2-23a)



- Questa tecnica è nota come **striping**, e in realtà era già adottata prima dell'introduzione del concetto di RAID.
- I produttori di sistemi RAID offrono, tra le configurazioni possibili, anche il livello 0, nel caso in cui l'utilizzatore voglia solo massimizzare le prestazioni e la capacità del sistema.

RAID livello 0 (block striping)

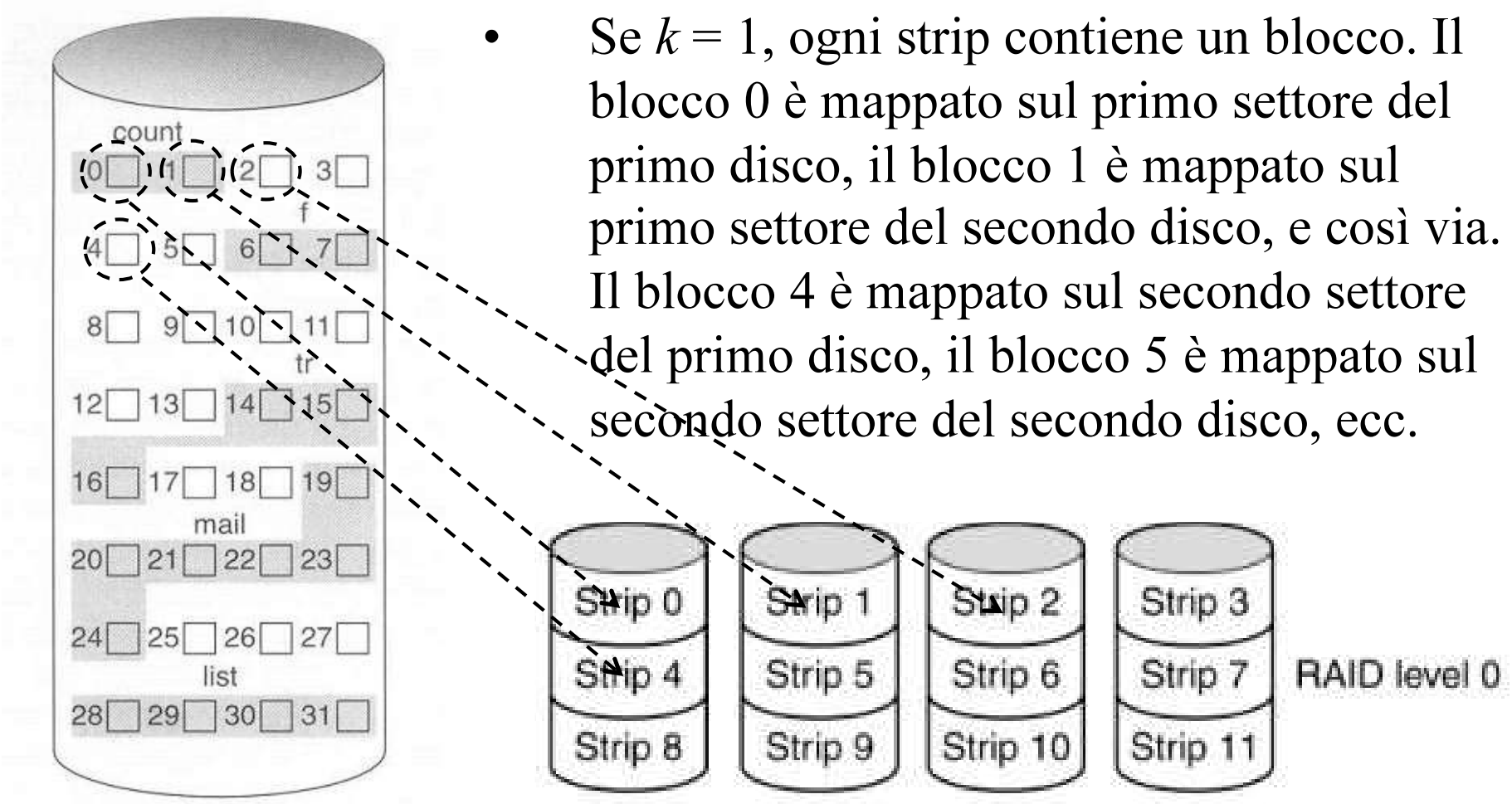
- Ma perché le prestazioni di un sistema RAID di livello 0, attraverso l'uso dello striping, sono migliori di un normale disco che memorizza le stesse informazioni?
- Supponiamo che il sistema operativo richieda la lettura (o scrittura) di un insieme di dati contenuti in 4 strip consecutivi. Ad esempio, i dati sono contenuti negli strip 4, 5, 6 e 7.
- Il controller RAID suddividerà la richiesta in 4 letture (scritture) che possono essere eseguite in parallelo sui 4 dischi del sistema, con un evidente aumento delle prestazioni rispetto alla lettura (scrittura) degli stessi settori su di un singolo disco, che deve leggere (scrivere) i settori di cui sono composti tutti gli strip in sequenza.
- Il SO non si accorge di nulla, se non che riceve i dati richiesti più velocemente che nel caso di un normale disco SLED

RAID livello 0 (block striping)

- Un RAID di livello 0 è tanto più efficiente quanto più le richieste coinvolgono l'accesso a molti blocchi consecutivi e quanto più è alto il numero di dischi su cui sono suddivisi i blocchi.
- Al contrario, operazioni su disco che richiedono l'accesso a pochi dati, ad esempio contenuti in un unico strip, non ottengono alcun miglioramento di prestazioni rispetto ad un disco SLED.
- inoltre, l'affidabilità di un sistema RAID di livello 0 è inferiore a quella di un semplice disco SLED, perché il RAID è formato di più dischi e il **Mean Time To Failure (MTTF)** non può che essere inferiore.
- Il RAID livello 0 viene usato in quelle applicazioni in cui sono necessarie alte prestazioni, senza particolari problemi di affidabilità (e.g., audio e video streaming)

Sistema Operativo e RAID

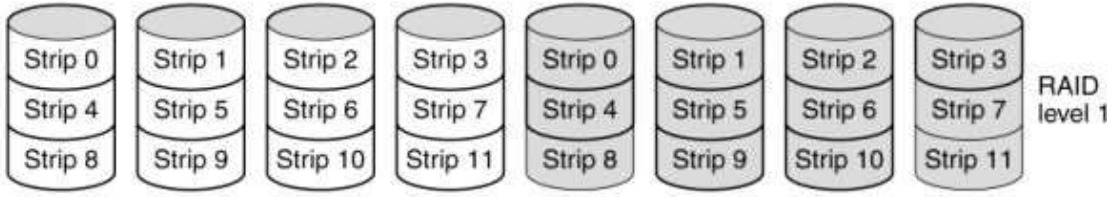
- Quindi, riassumento, il SO vede un disco virtuale con blocchi numerati consecutivamente che vengono suddivisi in strip e mappati sui vari dischi del RAID dal RAID controller



- Se $k = 1$, ogni strip contiene un blocco. Il blocco 0 è mappato sul primo settore del primo disco, il blocco 1 è mappato sul primo settore del secondo disco, e così via. Il blocco 4 è mappato sul secondo settore del primo disco, il blocco 5 è mappato sul secondo settore del secondo disco, ecc.

RAID livello 1 (striping+mirroring)

- Un RAID di livello 1 usa contemporaneamente striping e **mirroring**: tutti i dati sono suddivisi in strip (gestiti come nel livello 0) e duplicati su due dischi
- Quando un disco si rompe, il sistema di controllo del RAID si rivolge al disco di mirror per l'accesso ai dati. Nel frattempo, il disco rotto può essere sostituito (anche con procedure automatiche se è disponibile un ulteriore *spare disk,* su cui verranno copiati i dati contenuti nel "gemello" del disco rotto). (Tanenbaum, Fig. 2-23b)



RAID livello 1 (striping+mirroring)

- Il livello 1 è la soluzione RAID più costosa, a parità di capacità di memorizzazione, perché richiede la duplicazione di tutti i dischi.
- E' anche la soluzione più affidabile rispetto ai guasti, e la più efficiente in lettura: anche la lettura di un blocco di dati che coinvolge 5 strip (nel nostro esempio), può essere eseguita con cinque letture in parallelo, usando anche i dischi di mirroring.
- Il RAID livello 1 viene usato dove l'affidabilità è fondamentale, ad esempio per memorizzare dati finanziari e bancari

RAID livello 2
(bit level striping + ECC)

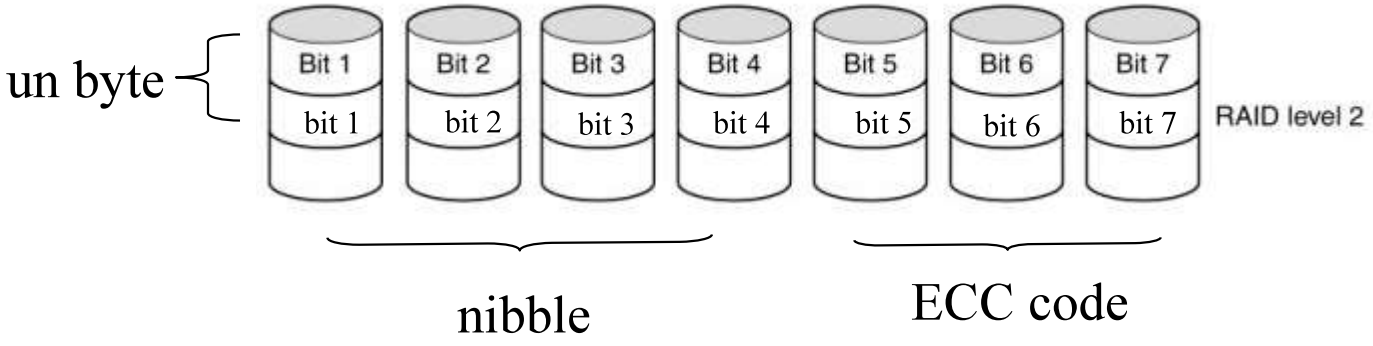
- Il RAID di livello 2 non è usato nei sistemi commerciali, ed è definito sostanzialmente solo a livello teorico: la sua implementazione richiede infatti molto overhead computazionale e l'uso di molti dischi che devono sempre essere sincronizzati in rotazione e nella posizione delle testine di lettura/scrittura.
- nel RAID livello 2 i dati vengono distribuiti tra i vari dischi addirittura al livello dei singoli bit che compongono una parola, insieme con un ECC (Error Correcting Code) per quei dati (ad esempio un codice di Hamming).
- (ricordate che il codice di Hamming associato ad gruppo di bit permette di correggere un bit errato nel gruppo)

RAID livello 2
(bit level striping + ECC)

- Consideriamo ad esempio word di 32 bit a cui siano aggiunti il codice di Hamming su 6 bit, e il bit di parità. Ciascuna word viene suddivisa su 32+6+1 dischi: lo stesso bit dello stesso settore di ciascun disco ospiterà uno dei 39 bit della word.
- I bit della word successiva saranno memorizzati a fianco di quelli della precedente sui 39 dischi, e così via.
- Se durante la lettura di dati un disco si rompe, significa che dei 39 bit di ciascuna word se ne è perso uno, che può essere ricostruito *immediatamente* sulla base degli altri usando il corrispondente ECC.

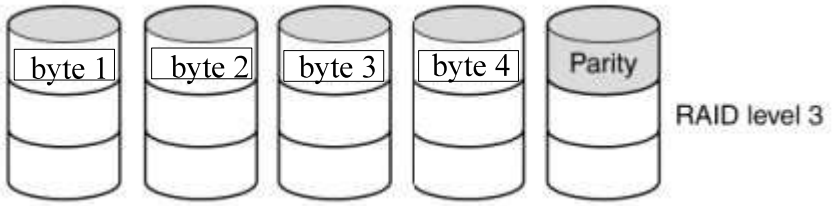
RAID livello 2
(bit level striping + ECC)

- Le prestazioni di questa configurazione sono (sarebbero) particolarmente elevate nel caso di lettura sequenziale dei dati: poiché i dischi ruotano in modo sincronizzato, la lettura contemporanea di un unico settore su tutti i dischi permette di leggere in realtà l'equivalente di 512 x 8 word consecutive da 32 bit (correggendo immediatamente eventuali errori singoli)
- Tanenbaum, Fig. 2-23c: il numero di dischi necessari per memorizzare un nibble (4bit) + il suo ECC.



RAID livello 3
(byte level striping + parity)

- Il RAID livello 3 è simile al RAID livello 2, ma lo striping avviene a livello dei byte di ciascuna parola, e viene calcolato il byte di parità per i byte memorizzati nella stessa posizione sui vari dischi. Questo livello RAID è disponibile nei sistemi commerciali, sebbene non sia usato molto spesso.
- La lettura sequenziale è molto efficiente anche in questo livello, che è particolarmente adatto per applicazioni (ad esempio in campo multimediale) che richiedono un accesso sequenziale veloce a file di grandi dimensioni, con anche un certo grado di affidabilità. (Tanenbaum, Fig. 2-23d)



RAID livello 3
(byte level striping + parity)

- Notate: il calcolo del byte di parità, per l'eventuale correzione degli errori, può essere fatto (ad esempio) calcolando l'EX-OR dei byte corrispondenti. Così,
  parità = byte 0 EX-OR byte 1 EX-OR byte 2 EX-OR byte 3.

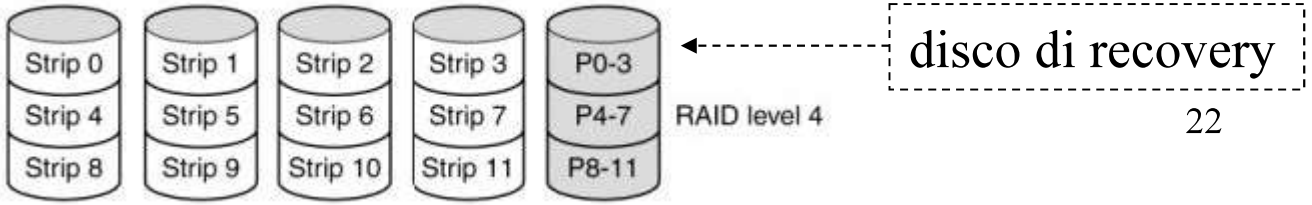
- Se un disco si rompe, è possibile ricostruire ogni byte che conteneva *sottraendo* al byte di parità i byte memorizzati nella stessa posizione negli altri dischi (in realtà basta rifare l'EX-OR).

Ad esempio, per tre byte:
a) 01100011 EX-OR
b) 10101010 EX-OR
c) 11001010 =
p) 00000011

e quindi, se si perde a):
p) 00000011 EX-OR
b) 10101010 EX-OR
c) 11001010 =
a) 01100011

RAID livello 4
(Block level striping with parity)

- Il RAID livello 4 è simile al livello 3 (con cui a volte viene confuso) ma usa la tecnica di striping a livello di blocchi, (come nei RAID livello 0 e 1), calcolandone la parità (come nel livello 3) per l'eventuale operazione di recovery.
- Un disco viene quindi usato per memorizzare la parità calcolata per gli strip che stanno nella stessa posizione negli altri dischi.
- Ad esempio, (Tanenbaum, Fig. 2-23e), nel caso di un sistema con 5 dischi, il primo strip del disco di recovery conterrà la parità calcolata usando gli strip 0, 1, 2 e 3, il secondo strip di parità conterrà la parità degli strip 4, 5, 6 e 7, e così via.

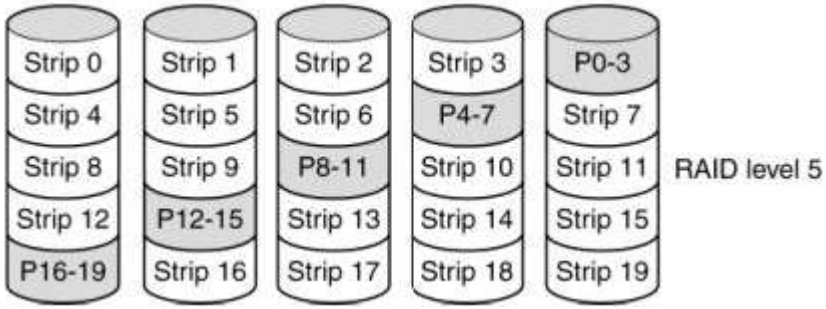


RAID livello 4
(Block level striping with parity)

- Il RAID livello 4 risparmia ovviamente dischi rispetto al RAID livello 1, e garantisce lo stesso il mantenimento dei dati in caso di guasto di un disco, ma al costo di una maggiore inefficienza.
- Infatti, ogni qualvolta uno strip di un disco viene modificato, occorre leggere anche i corrispondenti strip di tutti gli altri dischi e ricalcolarne la parità. Oppure, si può leggere il vecchio strip, sottrarlo allo strip di parità, e ricalcolare lo strip di parità usando il nuovo strip (notate che ciò è vero anche per il livello 3).
- Inoltre, il disco che ospita gli strip di parità è pesantemente coinvolto in ogni operazione di scrittura sul RAID, e può facilmente diventare un collo di bottiglia (il che vale anche per il livello 3).

RAID livello 5 (Block level striping with distributed parity)

- Il RAID livello 5 funziona sostanzialmente come il 4, ma per ridurre il carico sul disco di parità nelle operazioni di scrittura, il livello 5 distribuisce gli strip di parità fra i vari dischi (Tanenbaum, Fig. 2-23f)



- Il difetto di questo approccio è che, in caso di guasto di un disco, è più complessa la ricostruzione del suo contenuto, che è formato sia da strip di dati che da strip di parità
- Questo livello fornisce comunque la migliore combinazione in termini di prestazioni, affidabilità, e capacità di memorizzazione, ed è quindi di gran lunga il livello più usato per applicazioni generiche.

RAID livello 6 (Block level striping + dual distributed parity)

- (Attenzione: alcune compagnie usano “livello 6” per indicare in realtà loro prodotti specifici che sono estensioni del livello 5)
- L’ultimo livello RAID è il livello 6, in grado di resistere anche al guasto di due dischi contemporaneamente, combinando due livelli di parità, distribuiti sui vari dischi come nel caso del livello 5.
- Per questo sistema è necessario un disco in più del livello 5 per memorizzare la stessa quantità di dati, e un maggiore overhead computazionale.
- L’affidabilità di questo livello è ovviamente superiore a quella dei precedenti, ma questa soluzione è raramente usata perché la rottura contemporanea di due dischi è un evento estremamente raro (l’MTTF di un singolo disco è ormai stimato in anni)

RAID livello 7

- Esiste anche un RAID livello 7, che però non fa parte della definizione originale dello standard RAID, ed è in realtà è un nome registrato da una compagnia che produce sistemi di memorizzazione, la *Storage Computer Corporation*.
- Questo sistema RAID è una combinazione dei livelli 3 e 4, e offre prestazioni (in termini di velocità di accesso e affidabilità) superiori agli altri livelli RAID, ma anche a costi decisamente superiori.