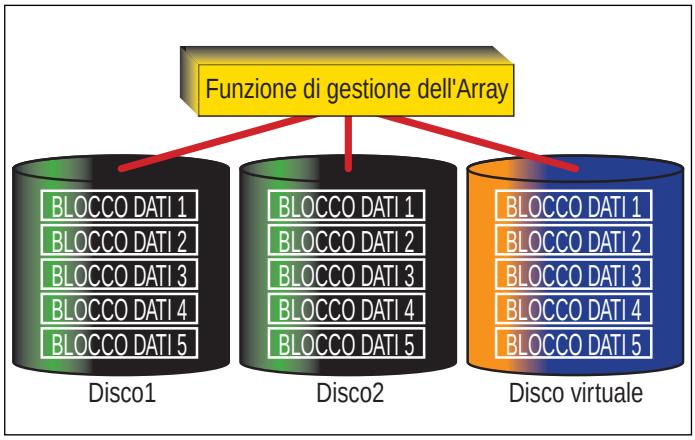
▶ Affidabilità dei dati

Introduzione al Raid

Nata per necessità quando i dischi erano costosi e poco affidabili, la tecnologia RAID oggi è alla portata di tutti, anche se non ha la diffusione che merita

Il termine RAID nacque nel 1987 per opera di Patterson, Gibson e Katz, tre ricercatori dell'università di California a Berkeley che scrissero un documento dal titolo *A case for Redundant Arrays of Inexpensive Disks* (RAID). Un *disk array* è una schiera, serie o gruppo di dischi; un array ridondante prevede una qualche forma di ripetizione dei dati, in modo da ricostruire il contenuto di qualsiasi disco dell'array che andasse fuori uso. 15 anni fa un'unità dischi da 7.500 MB costava 100.000 dollari, era voluminosa e dissipava parecchia energia. I ricercatori americani constatarono che gli economici dischi da 100 MB per PC avevano prestazioni non molto inferiori, mentre offrivano vantaggi in termini di costo per megabyte, ingombro e consumi. Una serie di dischi di basso costo, opportunamente raggruppati, si presentava come possibile alternativa all'uso dei costosi dischi ad alta capacità. Nella pubblicazione citatata gli autori proposero diverse soluzioni per migliorare le prestazioni e l'affidabilità dei disk array: modalità di accesso simultaneo ai dischi per ridurre i tempi delle operazioni di I/O e varie forme di ridondanza per poter ricostruire i contenuti dopo la sostituzione di un disco in avaria. A quel tempo occorrevano decine di dischi di basso costo per sostituire un singolo grande disco; considerando che il tempo medio prima di un guasto (MTTF – *Mean Time To Failure*) di un array è dato dall'MTTF di un singolo disco diviso per il numero di dischi dell'array, si vede come l'affidabilità fosse il problema principale da risolvere. Con le 30.000 ore di MTTF dei dischi da 100 MByte dell'87, un array con 100 dischi si sarebbe guastato mediamente ogni 12 giorni. Fortunatamente l'affidabilità dei dischi è cresciuta rapidamente nel tempo; da parecchi anni i migliori dischi SCSI in commercio presentano un MTBF (*Mean Time Between Failure*, tempo medio tra guasti) di un milione di ore. L'idea dei disk array con ridondanza, visti i grossi risparmi che prometteva, fu accolta con grande interesse e diede il via allo sviluppo di prodotti (chip, controller e hard disk) che oggi coprono tutte le fasce di prezzo e prestazioni. Negli anni '90 la differenza di prezzo tra i dischi per grossi computer e i migliori dischi per PC ha finito per scomparire, vista la rapida crescita della densità e capacità di registrazione, l'introduzione di testine e di tecnologie di lettura sempre più sofisticate e la crescente richiesta di alte capacità di archiviazione. Per questo il *RAID Advisory Board*, l'ente consultivo che raccoglie i maggiori produttori in questo campo, ha decretato la trasformazione dell'acronimo RAID in *Redundant Array of Independent Disks*. Le motivazioni che portano all'adozione di un sistema RAID sono essenzialmente tre: raggiungere alte capacità di archiviazione superando i limiti dei singoli dischi, ottenere prestazioni superiori e salvaguardare i dati e la continuità del servizio anche in caso di sostituzione di un disco guasto. Inoltre un disk array risolve automaticamente il problema del bilanciamento del carico tra più hard disk in termini sia di velocità di accesso sia di utilizzo dello spazio disponibile. Tra le esigenze citate, la capacità di archiviazione è risolta schierando il numero necessario di dischi e sfruttando la banda passante sempre più ampia messa a disposizione dai bus e dai controller. Le alte prestazioni si ottengono scegliendo la struttura RAID più adatta alle applicazioni e impiegando dischi veloci, bus ad alta banda passante e controller dotati di cache e di funzioni di ottimizzazione. L'accessibilità del sistema è garantita dalla ridondanza dei dati e dalla ricostruzione automatica dell'array tramite dischi di riserva mantenuti on-line; i contenuti del disco guasto (dati e informazioni di controllo) vengono ricalcolati e scritti sul primo disco libero a disposizione.

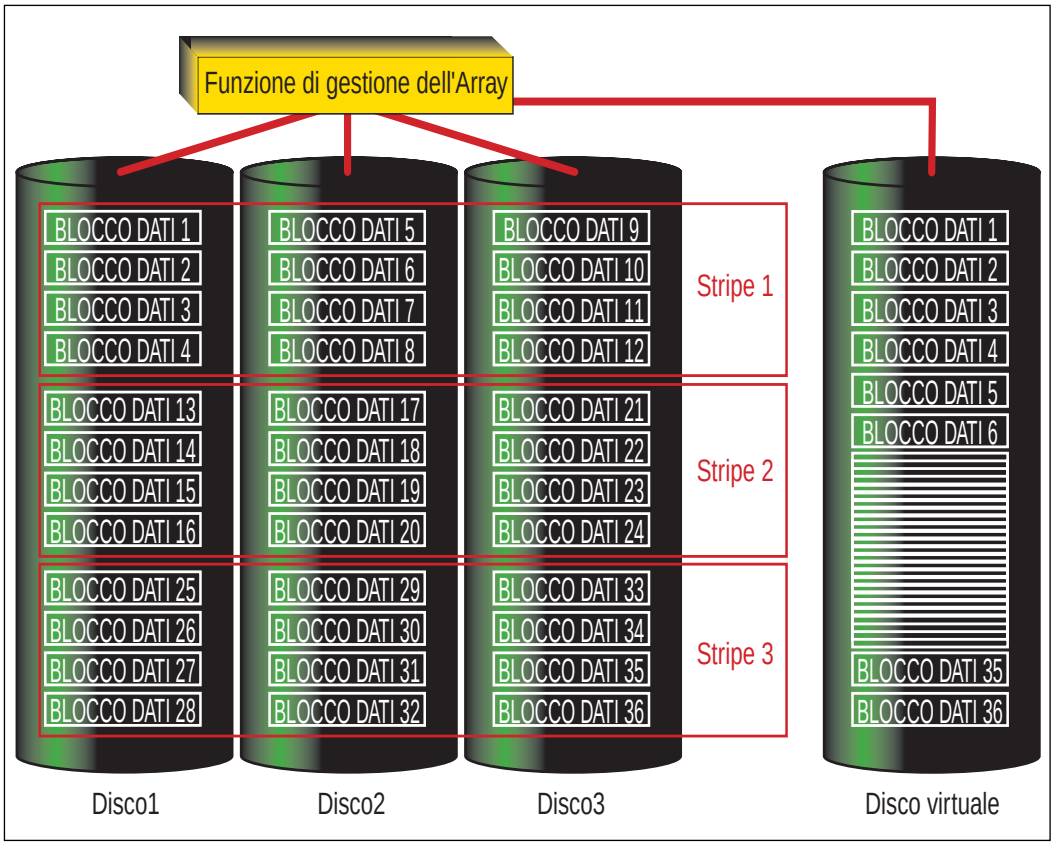
Nella Fotografia a sinistra *un controller RAID SCSI.* ***Sopra*** *il principale libro sul RAID*



In un array con **mirroring** *(RAID 1) i dischi 1 e 2 hanno lo stesso contenuto e sono visti come un singolo disco virtuale*

I livelli RAID

Nel documento originario del 1987 gli autori proposero cinque modi per creare gruppi ridondanti di dischi, assegnando un numero a ciascuno di essi. I vari livelli hanno i loro pro e contro e si prestano a campi di utilizzo diversi. Uno dei criteri fondamentali per scegliere quale livello RAID adottare è il tipo di applicazione: a seconda che si elaborino grandi quantità di dati sequenziali o numerose brevi transazioni simultanee si sceglierà una struttura che privilegi l'accesso parallelo o l'accesso indipendente ai dischi. La tabella che pubblichiamo, tratta dal RAID book (pubblicazione ufficiale del RAID Advisory Board e principale testo di riferimento in materia), riassume le caratteristiche essenziali sia dei cinque livelli RAID originari sia di due livelli aggiunti: il RAID 6, introdotto



RAID 0: è un array con striping e senza ridondanza. *I dati visti in sequenza dal sistema operativo sono distribuiti attraverso i dischi dell'array. Fra i vantaggi segnaliamo la capacità di trasferire grandi quantità di dati in input/output*

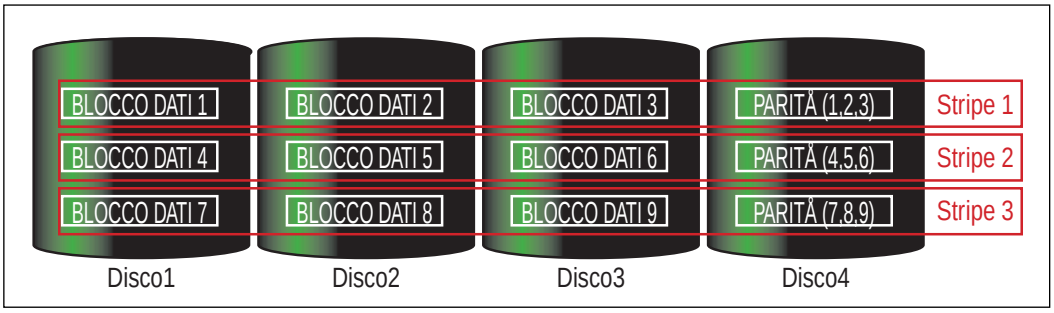
nel 1989 dagli stessi autori della pubblicazione originaria e il *RAID 0*, un array chiamato RAID per comodità ma privo di ridondanza e quindi privo di ogni misura di protezione in caso di guasto.

Perché una serie di dischi possa essere chiamata *disk array* deve essere presente, in hardware o in software, una funzione di gestione dell'array che governi il funzionamento dei dischi e ne presenti lo spazio di archiviazione al *computer host* sotto forma di uno o più dischi virtuali. Quasi sempre la gestione di un array è svolta da un *controller hardware* con l'ausilio del relativo *firmware* (comprendente utility di configurazione e diagnosi) e software (driver e utility di monitoraggio). Ad esempio, nel caso di due dischi in *mirroring* (RAID 1), dove ciascuno è specchio dell'altro, la funzione di *array management* provvede ad aggiornare entrambi i dischi ogni volta che viene eseguita una scrittura. Se il controller RAID fosse abbastanza sofisticato, la velocità di lettura nel RAID 1 potrebbe essere sensibilmente superiore a quella con singolo disco, perché si potrebbe leg- ▷

I LIVELLI RAID STANDARD

Livello RAID	Nome comune	Descrizione	Dischi richiesti	Disponibilità dei dati	Capacità di trasferimento di grandi quantità di I/O	Capacità di eseguire richieste di piccoli I/O
0	Disk striping	Dati distribuiti attraverso i dischi dell'array. Nessun dato di verifica	N	inferiore rispetto a un solo disco	molto alta	molto alta in lettura e in scrittura
1	Mirroring	Tutti i dati sono replicati su N dischi separati (N di solito è uguale a 2)	2N, 3N ecc.	maggiore che nei RAID di Livello 2, 3, 4, 5; inferiore rispetto al RAID di Livello 6	maggiore che in un disco singolo in lettura; simile a un disco singolo in scrittura	fino a due volte quella di un disco singolo in lettura; simile a un disco singolo in scrittura
2	n.d.	Dati protetti con codice Hamming. Dati di verifica distribuiti su m dischi, dove m è determinato dal numero di dischi di dati nell'array	N+m	molto più alta che con un disco singolo; più alta che con RAID 3, 4 o 5	la più alta tra le alternative elencate	circa il doppio rispetto a un disco singolo
3	RAID 3, Dischi a trasferimento parallelo con parità	Ogni blocco di disco virtuale è suddiviso e distribuito attraverso tutti i dischi di dati. I dati di verifica di parità sono archiviati su un disco di parità separato.	N+1	molto più alta che con un disco singolo; paragonabile a RAID 2, 4 o 5	la più alta tra le alternative elencate	circa il doppio rispetto a un disco singolo
4	n.d.	Blocchi di dati distribuiti come nel disk striping. Dati di verifica di parità archiviati su un disco	N+1	molto più alta che con un disco singolo; paragonabile a RAID 2, 3 o 5	simile a disk striping in lettura; notevolmente inferiore a un disco singolo in scrittura	simile a disk striping in lettura; notevolmente inferiore a un disco singolo in scrittura
5	RAID 5, o *RAID*	Blocchi di dati distribuiti come nel disk striping. Dati di verifica di parità distribuiti su più dischi	N+1	molto più alta che con un disco singolo; paragonabile a RAID 2, 3 o 4	simile a disk striping in lettura; inferiore a un disco singolo in scrittura	simile a disk striping in lettura; generalmente inferiore a un disco singolo in scrittura
6	RAID 6	Come per RAID 5, con l'aggiunta di dati di verifica calcolati in modo indipendente	N+2	la più alta tra le alternative elencate	simile a disk striping in lettura; inferiore a RAID 5 in scrittura	simile a disk striping in lettura; notevolmente inferiore a RAID 5 in scrittura

gere contemporaneamente dai due dischi. Tranne che nel caso del mirroring, dove si ha la replica dell'intero contenuto di uno o più dischi, un array implica la distribuzione dei dati su più dischi attraverso lo *striping*, cioè la distribuzione a ventaglio dei dati sui dischi che formano l'array. Ad esempio nel RAID 0, quando si registra un file, un certo numero di blocchi viene registrato in sequenza sul primo disco dell'array, altrettanti sul secondo disco e così via fino all'ultimo, per poi ricominciare dal primo distribuendo a rotazione i dati sui vari dischi. In questo modo si realizza una sovrapposizione temporale delle operazioni di I/O sui dischi che accelera notevolmente le prestazioni nel caso di grossi trasferimenti sequenziali (una cache sul controller accelera ulteriormente l'I/O su uno striped array). Nei livelli RAID da 2 a 6 gli array sono caratterizzati da varie forme di striping e dalla presenza di ridondanza sotto forma di *check information*, cioè informazioni di verifica che permettono di ricostruire i dati di qualunque disco che andasse in avaria. Resta inteso che la ricostruzione è possibile se si guasta un disco alla volta; in ogni caso, visto che esistono altre cause di perdita di dati oltre ai guasti dei dischi, la presenza del RAID non elimina la necessità di una strategia di backup



Raid 3: *un array ad accesso parallelo con parirà su un solo disco. I dati sono distribuiti byte per byte attraverso i dischi dell'array*

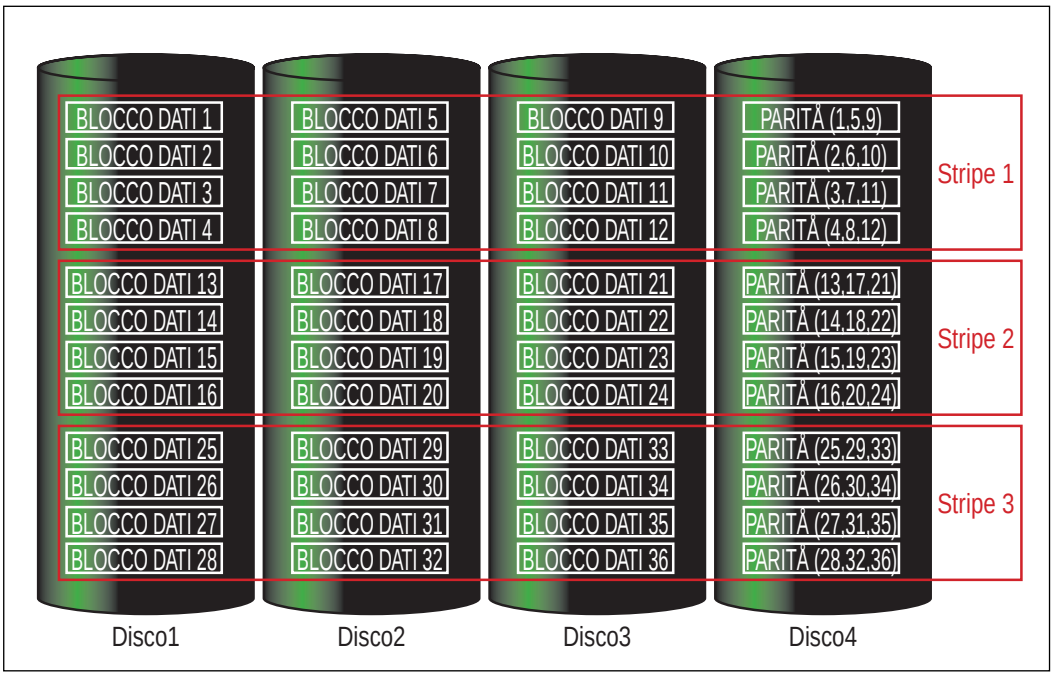
e l'opportunità di adottare misure di *duplexing* (raddoppio dei componenti) e *fault tolerance* sul sistema (in particolare per alimentatori, ventole, controller e cache).

Nel RAID 2 la ridondanza proposta dagli autori di Berkeley doveva imitare la correzione di errori ECC adottata nelle memorie RAM di server e workstation, dove un gruppo aggiuntivo di bit permette di correggere singoli bit errati e di segnalare errori su due bit. I dati vengono distribuiti bit per bit (byte per byte in pratica) attraverso i dischi dell'array, con l'aggiunta di un numero sufficiente di *check disk* (dischi con i bit di correzione) per correggere singoli bit errati nei dati. Per un array con 10 dischi di dati occorrono quattro dischi di verifica aggiuntivi. Le prestazioni sono buone soprattutto per i trasferimenti di grandi quantità di dati e accesso parallelo; il RAID 2 non si presta per gli accessi indipendenti di un ambiente transazionale. Dato che l'unità minima di trasferimento è di 512 byte (un settore), la dimensione dei dati trasferiti deve essere di almeno 512 byte per il numero di dischi destinati ai dati. Questa forma di RAID non ha avuto attuazione pratica anche perché avrebbe richiesto modifiche al progetto dei dischi.

Il raid 3 e 4

Il RAID 3 è simile al RAID 2 perché utilizza lo stesso striping byte per byte (con tre dischi per i dati avremmo il byte 0 sul primo disco, il byte 1 sul secondo disco, il byte 2 sul terzo, il byte 3 sul primo e così via). La differenza sta nel fatto che le informazioni di verifica (bit di parità) sono concentrate su un unico disco aggiuntivo. I dati di un disco in avaria vengono ricostruiti calcolando la parità dei dischi rimasti operativi e confrontandola bit per bit con quella del disco di verifica; quando le parità concordano, il bit da ricostruire è 0, altrimenti è 1. Le prestazioni del RAID 3 sono simili a quelle del RAID 2, salvo che occorrono meno dischi e quindi è più favorevole il calcolo delle prestazioni per disco, uno dei parametri nel confronto tra i livelli RAID. Anche il RAID 3 è ad accesso parallelo, adatto ai supercomputer e alla manipolazione di grossi file (video e così via) ma non alle transazioni indipendenti (*multithreaded*) di un database o di un server (nei RAID 2 e 3 ogni operazione di I/O utilizza tutti i dischi). Inoltre lo striping dei byte sui dischi richiede il sincronismo nella rotazione dei dischi dell'array (*spindle sync*, sincronismo degli alberi dei dischi) per evitare i tempi di attesa. Il documento originale cita un fattore *S* (*slowdown*, rallentamento), compreso tra 1 e 2, per il quale va diviso il numero previsto di I/O al secondo, dovuto all'attesa che tutti i dischi di un array finiscano di leggere o scrivere un settore. Se i dischi vengono sincronizzati, non c'è rallentamento e *S* vale 1. Nel calcolo delle prestazioni il fattore *S* è presente per tutti i livelli RAID tranne per i brevi I/O nei RAID 1, 4 e 5, che avvengono in modo indipendente. Prove di RAID 3 con gli attuali hard disk, che non supportano più il sincronismo, hanno mostrato prestazioni deludenti.

Il RAID 4 è il primo livello di *striped array* adatto a impieghi transazionali. Anche in questo caso i dati sono distribuiti in stripe attraverso i dischi, salvo che, anziché trasferire un bit o un byte per disco, vengono trasferiti abbastanza byte da alloggiare l'intero record della transazione. Si perde il vantaggio velocistico di distribuire i



RAID 4: *un array ad accesso indipendente con parità su un solo disco*

dati simultaneamente sui dischi ma si guadagna la possibilità di realizzare diversi brevi I/O simultanei sui dischi dell'array. Come nel RAID 3, i dati di parità risiedono su un unico disco separato dai dischi dei dati. Poiché ogni settore di dati risiede su un solo disco, anziché essere distribuito byte per byte sui dischi dell'array, il calcolo della parità è molto più semplice. Anche se nel RAID 4 una breve operazione di lettura richiede solo l'accesso a un disco, una altrettanto breve scrittura richiede quattro accessi: due letture e due scritture, quanto basta per rendere dubbia la sua utilità negli ambienti di transaction processing (magari a favore del RAID 1).

Inoltre, sebbene il RAID 4 sfrutti il parallelismo delle operazioni di lettura, le scritture sono limitate a una alla volta, perché ogni scrittura richiede la lettura e scrittura del disco di parità. Il RAID 5 risolve questo problema distribuendo per settori sui dischi dell'array sia i dati sia le informazioni di verifica, un piccolo cambiamento che ha un grosso impatto sulle prestazioni, dato che elimina il collo di bottiglia in scrittura e permette di eseguire scritture multiple simultanee. Dal momento che i moderni dischi SCSI ad alte prestazioni non supportano più il sincronismo di rotazione, il RAID 3 risulta pressoché inutilizzabile, quindi si usano combinazioni di RAID 0 e 1 e il RAID 5, diventato standard nei server e sempre veloce in lettura.

Il RAID 6, proposto dai ricercatori di Berkeley nel 1989, è un'evoluzione del RAID 5 che protegge l'array dall'avaria simultanea di due dischi tramite due forme indipendenti di informazioni di verifica. Sebbene il RAID 6 offra un'affidabilità eccezionale (anche sproporzionata rispetto alla probabilità di guasto degli altri componenti di un sistema), introduce un'ulteriore penalizzazione sulle operazioni di scrittura rispetto al RAID 5, perché le informazioni di verifica devono essere calcolate e scritte due volte a ogni aggiornamento. Il RAID 6 non è comunemente usato, anche perché l'affidabilità di un sistema RAID 5 oggi è del tutto adeguata.

Questi livelli RAID possono essere mescolati creando livelli ibridi; si possono avere *striped array di mirrored array* (array di tipo 0, 3 o 5 i cui componenti sono array RAID 1) o *mirrored array di striped array*, inferiori perché, a parità di prestazioni, sono più vulnerabili nella situazione di funzionamento degradato dopo l'avaria a un disco. Quindi il RAID 0+1 (*stripe of mirrors*) è preferibile rispetto alla soluzione 1+0 (*mirror of stripes*).

Quanto sopra è stato ricavato prendendo come riferimento le pubblicazioni citate, che sono le fonti ufficiali e più attendibili a cui ricorrere per cominciare a discutere di RAID. L'originario A Case for Redundant Arrays of Inexpensive Disks è diventato praticamente un cult-report, tante sono le volte in cui è stato citato nella letteratura. The RAID book rappresenta il punto di vista tecnico e asettico, al di sopra delle parti, del RAID Advisory Board, che fa opera educativa e di standardizzazione nel campo dell'affidabilità dei sistemi di archiviazione su disco. Dato che non basta avere un RAID per dormire sonni tranquilli, il RAID Advisory Board si occupa anche di EDAP (*Extended Data Availability and Protection*, disponibilità e protezione estesa dei dati) e in questo ambito ha introdotto una classificazione per i sistemi a disco (*Failure Resistant, Failure Tolerant* e *Disaster Tolerant*) che utilizza per certificare la conformità dei sistemi sottoposti a valutazione dai produttori. ■

Giorgio Gobbi

GLOSSARIO

ARRAY (DISK ARRAY)
È una raccolta di dischi di comune reperibilità combinata con una Funzione di Gestione dell'Array, che controlla le operazioni sui dischi e presenta la capacità di archiviazione globale sotto l'aspetto di uno o più dischi virtuali.

ARRAY AD ACCESSO INDIPENDENTE
Un disk array la cui mappatura è tale che i diversi dischi che compongono l'array possono eseguire contemporaneamente più richieste di I/O da parte dell'applicazione.

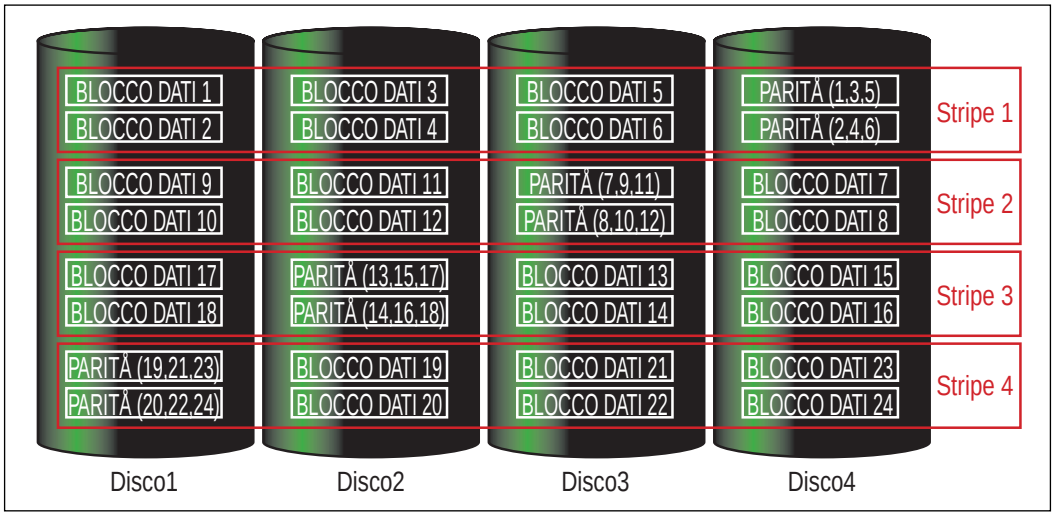
ARRAY AD ACCESSO PARALLELO
Un disk array in cui il modello di accesso ai dati presuppone che tutti i dischi dell'array operino all'unisono e partecipino tutti all'esecuzione di ogni richiesta di I/O da parte delle applicazioni. Un array ad accesso parallelo è intrinsecamente capace di eseguire una sola richiesta di I/O alla volta. Gli array veramente paralleli richiedono il sincronismo fisico dei dischi (spindle sync); molto più spesso gli array approssimano il vero comportamento parallelo.

CHUNK O STRIP
La parte di stripe registrata su ognuno dei dischi di un array.

RIDONDANZA
L'inclusione di componenti aggiuntivi in un sistema al di là di quelli richiesti per svolgere la sua funzione. Le informazioni di verifica (check information) in un RAID sono la componente ridondante e servono a ricostruire i dati perduti in seguito all'avaria di un disco. Nel RAID 1 le informazioni di verifica sono l'intera copia dei dati; nei livelli da 2 a 4 sono bit di parità residenti su uno o più dischi appositi, mentre nel RAID 5 sono bit di parità inframmezzati ai dati.

STRIPE
La "striscia" risultante dall'operazione di striping, ovvero la ripartizione sui dischi di un array di un segmento di dati che appare contiguo all'ambiente operativo. La dimensione della stripe viene scelta secondo il tipo di applicazione. Valori comuni sono 32 e 64 KByte.

STRIPING (letteralmente fare a strisce)
È l'abbreviazione di disk striping ed è sinonimo di RAID 0. È una tecnica di mappatura in cui segmenti consecutivi dei dati del disco virtuale (come l'ambiente operativo vede i dati di un array) sono mappati ai dischi dell'array in una sequenza ciclica. Una serie di dati che appare contigua all'ambiente operativo è in realtà distribuita parte sul primo disco dell'array, parte sul secondo, ... parte sull'ultimo e quindi da capo. Da solo, il disk striping non include alcuna protezione dei dati.



RAID 5: è la forma di RAID più usata nei server. *È un array ad accesso indipendente con le informazioni di parità distribuite su tutti i dischi dell'array*