

# Tecniche per la salvaguardia della disponibilità ed integrità dei sistemi di elaborazione e delle informazioni

## 4. backup

Marco Prandini  
Università di Bologna



## Disponibilità a medio e lungo termine

- Per quanto replicato, un sistema di memorizzazione dati non può
  - ignorare comandi espliciti, ma errati, di cancellazione/alterazione
    - Impartiti da utenti
    - Provocati da bug software (applicativo o di sistema)
  - sopravvivere ad eventi disastrosi
  - conservare un numero arbitrario di immagini della situazione fotografata ad un dato istante
- Tutte queste situazioni sono però di interesse pratico, e richiedono l'implementazione di
  - sistemi di *backup*
  - politiche di *recovery*



# Backup

- Il backup è la copia dei dati dal sistema live ad un supporto offline
  - è impegnativo organizzativamente e tecnicamente
  - è l'assicurazione contro qualsiasi causa di distruzione dei dati del sistema principale
- Va pianificato, considerando tra gli altri questi fattori:
  - cosa copiare (compromesso tra praticità di ripristino e tempi/spazi necessari)
  - chi è incaricato dei backup
  - quando è necessario/possibile eseguire il backup
  - quanto rapidamente cambiano i dati sul sistema
  - quanto velocemente deve poter essere eseguito il restore
  - per quanto deve essere conservata ogni copia
  - dove saranno conservate le copie
  - dove saranno ripristinate le copie (compatibilità cross-platform)

3

## Backup - strategie

- FULL BACKUP – è la copia completa di ogni singolo file nel/nei filesystem oggetto del backup
  - lento e ingombrante --> difficile farlo frequentemente
  - massima semplicità di ripristino
- INCREMENTAL BACKUP – è la copia dei soli file cambiati da una data di riferimento, tipicamente quella di esecuzione dell'ultimo full backup
  - adatto all'esecuzione frequente
    - attenzione al carico della “semplice” operazione di indicizzazione
  - per il ripristino servono sia il full che l'incremental
  - può essere realizzato anche a più livelli
    - Full
      - Incremental/level0/volume1 (rispetto al full)
        - incremental/level1/volume1 (rispetto all'incremental/0/1)
        - incremental/level1/volume2 (rispetto all'incremental/0/1)
      - Incremental/level0/volume2 (rispetto al full)

4

## Backup - cautele

- **Correttezza della copia** – idealmente il filesystem dovrebbe essere a riposo durante il backup, ma è raro nella pratica, quindi bisogna curare bene i dettagli relativi alla lettura di file aperti o di strutture complesse come i database
- **Protezione dei dati** – un backup contiene tutti i file del sistema, quindi in caso di requisiti di riservatezza va difeso allo stesso modo
- **Integrità dei dati** – se il backup viene svolto senza supervisione del sysadm, ci si deve cautelare da attività anche involontarie degli utenti che possano provocare la sovrascrittura dei dati
- **Affidabilità dei supporti** – con periodicità dipendente dalla criticità dei sistemi, ci si deve accertare che i dati siano scritti correttamente e siano leggibili per tutta la durata prevista della copia, curando
  - fattori tecnologici (graffi, smagnetizzazione, *obsolescenza hw e sw...*)
  - fattori ambientali (polvere, umidità, temperatura, ...)
- **Facilità di reperimento** – i supporti devono essere organizzati per consentire di individuare facilmente ciò che si deve ripristinare

5

## Backup – tecnologie

- **Tradizionalmente i backup venivano fatti su nastro già per sistemi di fascia medio-bassa**
  - basso costo per byte
  - alta capacità
  - **diverse soluzioni proprietarie ed incompatibili**
- **La crescita straordinaria della capacità degli hard disk ha messo in crisi le soluzioni tradizionali a nastro**
  - per sistemi di fascia bassa è comune l'approccio disk-to-disk
  - per sistemi di fascia alta sono state sviluppate soluzioni a nastro estremamente performanti e con un alto costo d'ingresso, compensato dal basso costo marginale (per GB)

6

# Backup – tecnologie

## ■ I supporti ottici sono poco utilizzati

- L'unico vantaggio è il basso costo del drive
- La massima capacità attualmente disponibile (blue-ray) è 100GB
  - 40 BD per 1 HD da 4TB
- La caratteristica di grande interesse dei supporti ottici è WORM (Write-Once Read-Many): i dati sono inalterabili una volta scritti
  - Affidabilità contro incidenti
  - Valore legale dell'archivio
  - ... ma anche i nastri più recenti la offrono
- Applicazioni di nicchia
  - Grandi data banks multimediali
    - Soluzioni Sony fino a 3.3TB
  - Esigenze di lunga conservazione
    - Shelf life oltre 100 anni

7

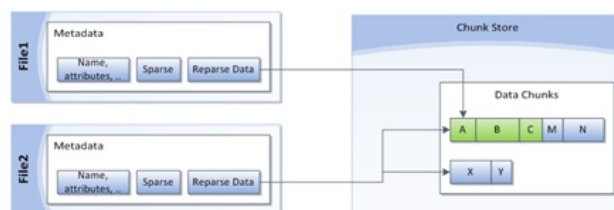
# Backup – tecnologie

## ■ Trend più recente: *data deduplication*

- Non solo per backup ma anche per main storage (es. ZFS)



- Dataset diviso in *chunk*, identificati da un hash → se un chunk ha lo stesso hash di un altro, viene eliminato e sostituito da un puntatore



- In teoria soffre del problema delle collisioni delle funzioni hash
- In pratica la probabilità di una collisione è enormemente più bassa di qualsiasi altro errore nella catena di storage

8

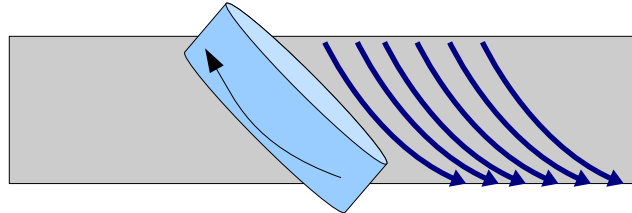
# Tecnologie per backup su nastro

## ■ Tipi di nastro

– Assunto di base: alto bitrate = alta velocità relativa tra nastro e testina

– Helical scan:

- La velocità è ottenuta per rotazione della testina → necessità di inclinazione per utilizzare “tracce elicoidali” diverse sul nastro

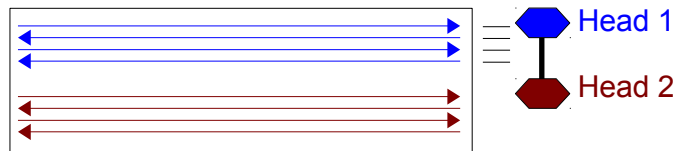


- Tipicamente testine magnetiche alternate sul tamburo scrivono / verificano / se necessario riscrivono / riverificano → complessità e fragilità meccanica

– Linear tape

- Singola testina fissa, il nastro scorre ad elevata velocità e viene scritto ad elevato bitrate → se non alimentato alla stessa velocità, accelera, vuota il buffer, decelera e poi arretra (*shoe shining*)

- La scrittura avviene “a serpentina” su tracce parallele per realizzare un nastro virtuale più lungo



9

# Tecnologie per backup su nastro a confronto

	Anno	Capacità (TB)	Velocità (MB/s)	Tipo	Vita		Costo (US\$)		
					anni	cicli	unità	nastro	c\$/GB
DDS-6 (DAT-160)	2007	0.08	7	Helical	10	2000	<<1k	30	37.50
DAT-320	2009	0.16	12	Helical	10	2000	1k	55	34.38
DLT-V4 (value)	2005	0.16	10	Linear	30		<<1k	40	25
DLT-S4 (hi perf)	2006	0.8	60	Linear	30		2k	80	10
LTO-5	2010	1.5	80-140	Linear	30	5000	<1k	15	1
LTO-6	2012	2.5	40-160	Linear	30	20k	1.5k	25	1
LTO-7	2015	6	40-300	Linear	30	20k	3k	75	1.25
LTO-8	2017	12	40-360	Linear	30		4k	180	1.5

Note:

- I prezzi sono in continua evoluzione al ribasso, soprattutto per le tecnologie più recenti – nella tabella sono aggiornati al 2018 solo per la tecnologia LTO (per i drive, è riportato un prezzo rappresentativo, ma la variabilità è molto alta)
- I cicli di utilizzo sono calcolati sull'uso incrementale (i.e. la “vita vera” è ad esempio 100 passate complete del nastro, si stima che ogni ciclo di utilizzo impegni 1/20 di nastro, da cui una vita stimata di 2000 cicli)
- Attenzione alla capacità *dichiarata*: spesso è 2 o 3 volte quella *reale* (riportata in tabella) perchè il venditore assume che sia mediamente raggiunto tale fattore di compressione (i drive comprimono in hardware)
- Per confronto:
  - Il costo per GB di un hard disk è tra 3 e 6 c\$/GB – la capacità raggiunge i 12TB
  - Il costo per GB di un BD-R è tra 2 e 5 c\$/GB – la capacità raggiunge i 100GB
- Caratteristiche uniche di LTO:
  - Ampio consorzio (*Linear Tape Open*)
  - Supporto HW all'automazione delle procedure di gestione dei nastri (i nastri hanno una memoria per i dati che li identificano, leggibile per via NFC)
  - Supporto HW WORM per garanzia di integrità e AES per riservatezza
  - LTFS (LTO-5 e superiori): possibilità di partizionare un nastro e formattare le partizioni come se fosse un disco

10



# Tecnologie per backup su nastro - librerie

## ■ I nastri sono vantaggiosi sui dischi

- per elevate capacità, dato il minor costo per GB
- per la lunga durata: una volta riempito, un volume può essere rimosso dal drive e conservato per decine di anni



Es: 1 o 2 drive,  
48 nastri LTO

Es: da 4 a 96 drive,  
da 350 a 6000 nastri

6000 nastri LTO4 = 4,5PB  
6000 nastri LTO6 = 15 PB  
(online, più la possibilità di archiviare facilmente i nastri che non servono con breve preavviso)

## ■ Per usarli in modalità simil-disco (tempi di accesso dell'ordine di 5-6 minuti) si può automatizzare il sistema di caricamento nel drive con una tape library



11

## Disaster recovery

- Tutte le misure discusse sono necessarie per limitare l'impatto degli inconvenienti pressochè quotidiani nell'uso dei calcolatori
- Per garantire la sopravvivenza di un'organizzazione ad eventi di portata catastrofica bisogna adottare precauzioni ulteriori
  - off-site storage – full backup attentamente verificato, conservato in luogo diverso da quello dei sistemi, anche in molteplici copie in luoghi diversi.
    - ripristino di un intero sistema a partire dalla macchina vergine (*bare metal*) – non solo dati ma partizionamento, struttura del filesystem, boot loader, sistema operativo, ...
  - site replication – realizzazione di un'intero duplicato dell'architettura di elaborazione dati dell'azienda, in luogo diverso dal sistema principale, e costantemente sincronizzato con questo
    - problematiche complesse di consistenza dei dati ed integrità dell'infrastruttura, a volte più gravi in failback che in failover

12

# Business continuity

## ■ Serve tutto questo?

- Delle imprese che hanno subito disastri con pesanti perdite di dati, circa il 43% non ha più ripreso l'attività, il 51% ha chiuso entro due anni e solo il 6% è riuscita a sopravvivere nel lungo termine. ([http://it.wikipedia.org/wiki/Disaster\\_recovery](http://it.wikipedia.org/wiki/Disaster_recovery))

## ■ Serve ma non basta

- HA, backup, DR sono **tecnologie**
- Un piano efficace di protezione dell'operatività aziendale (Business Continuity Plan) le prevede tutte, ma soprattutto ne definisce e documenta l'uso per mezzo di

### **PROCEDURE**

condivise, periodicamente aggiornate, periodicamente testate, possibilmente certificate a norma BS 25999-2

