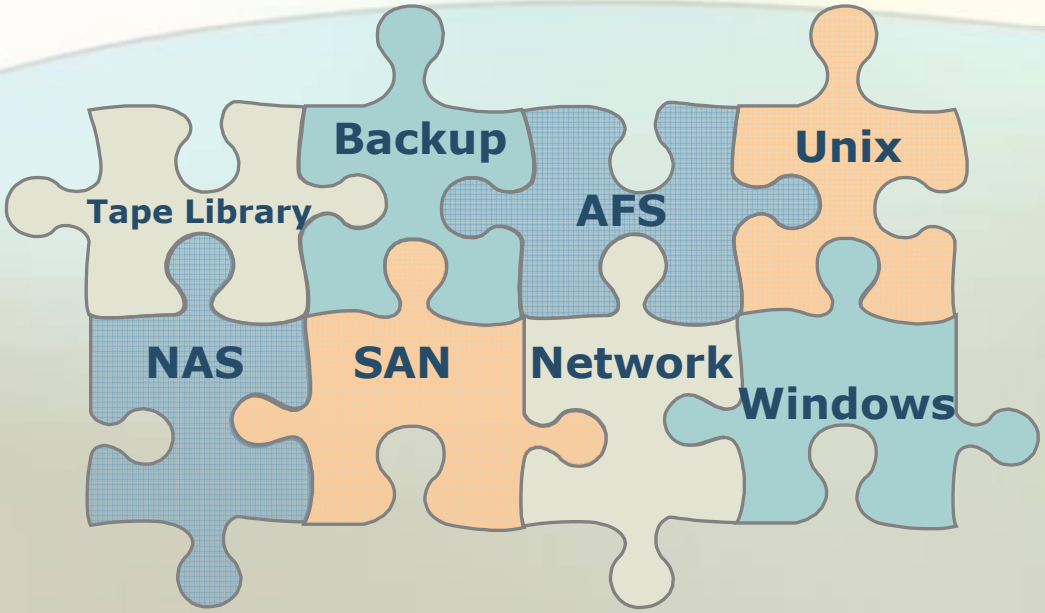
Memorie di Massa e File System

Sandro.Angius@lnf.infn.it



LNF – Settembre 2005

Argomenti Trattati

- Dischi
  - IDE, ATA, SATA, SCSI, SSA, Fiber Channel
  - JBOD, RAID, LVM
- File System
  - Locali (Ext2, Ext3, XFS)
  - Distribuiti (NFS, AFS)
- Nastri
  - Backup
  - Archiviazione
- Cenni sui Sistemi SAN, NAS
- Cenni su HSM, ILM

Cenni Storici

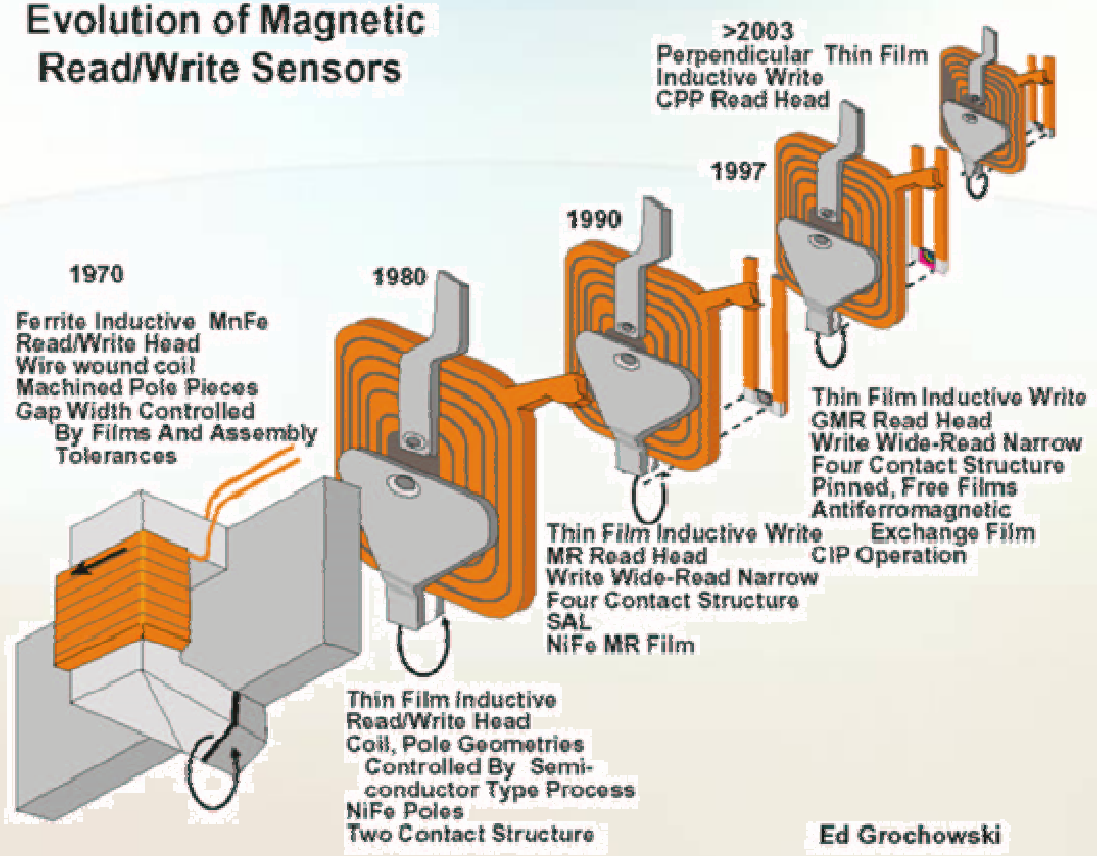
Il primo “hard disk” e’ stato prodotto nel 1956 dalla IBM.
Il “350 Disk Storage” era parte del sistema “IBM 305 RAMAC” (Random Access Memory Accounting).

Ogni disco girava a 1200 rpm,
Seek time medio ~600ms,
~5 Mbytes di capacita’.

Da allora le performance sono notevolemente aumentate ma il principio e’ rimasto lo stesso

HITACHI
Inspire the Next

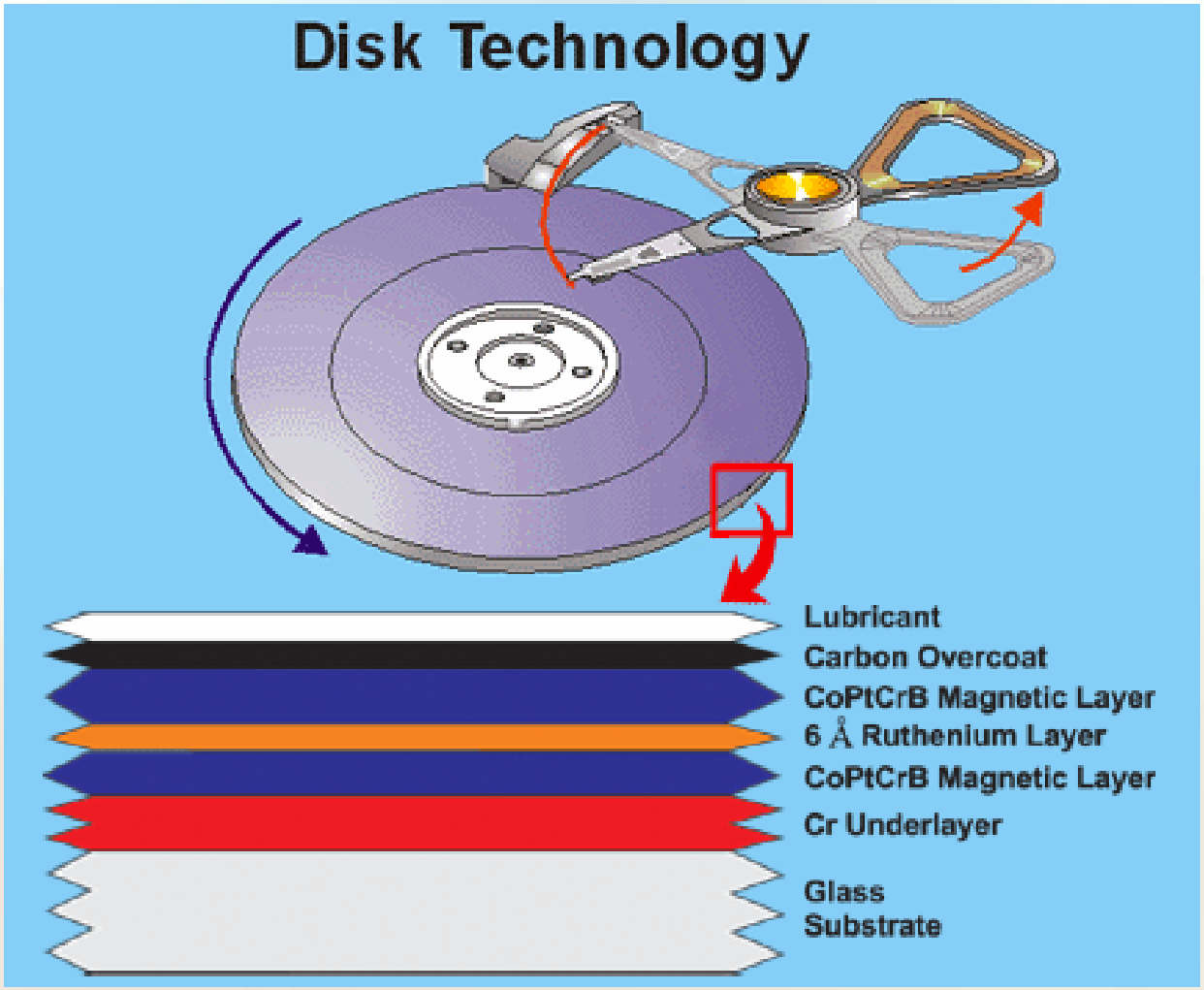
Evolution of Magnetic Read/Write Sensors



Ed Grochowski

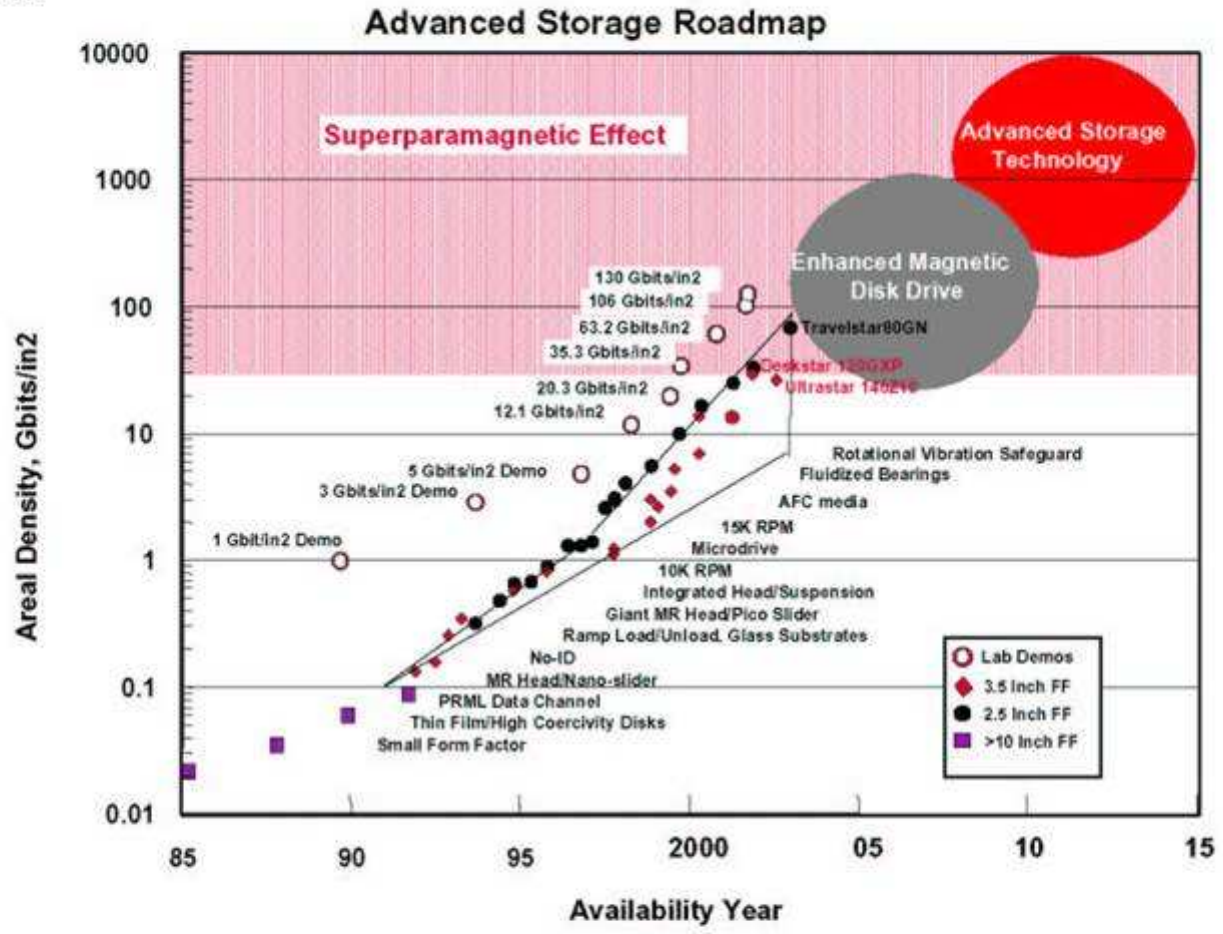
San Jose Research Center

Hitachi Global Storage Technologies



http://www.hitachigst.com/hdd/research/storage/adt/

HITACHI
Inspire the Next

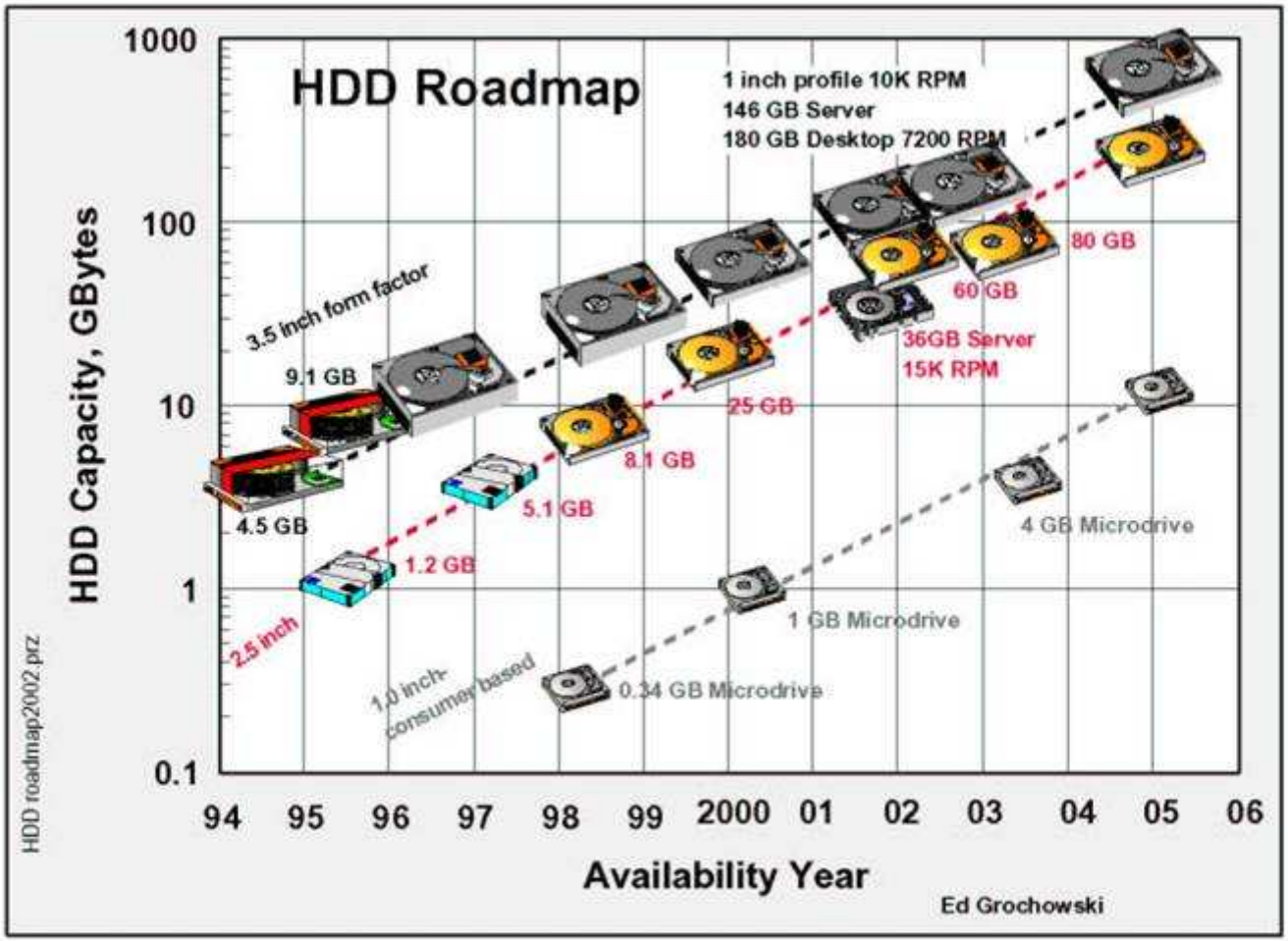


Ed Grochowski

San Jose Research Center

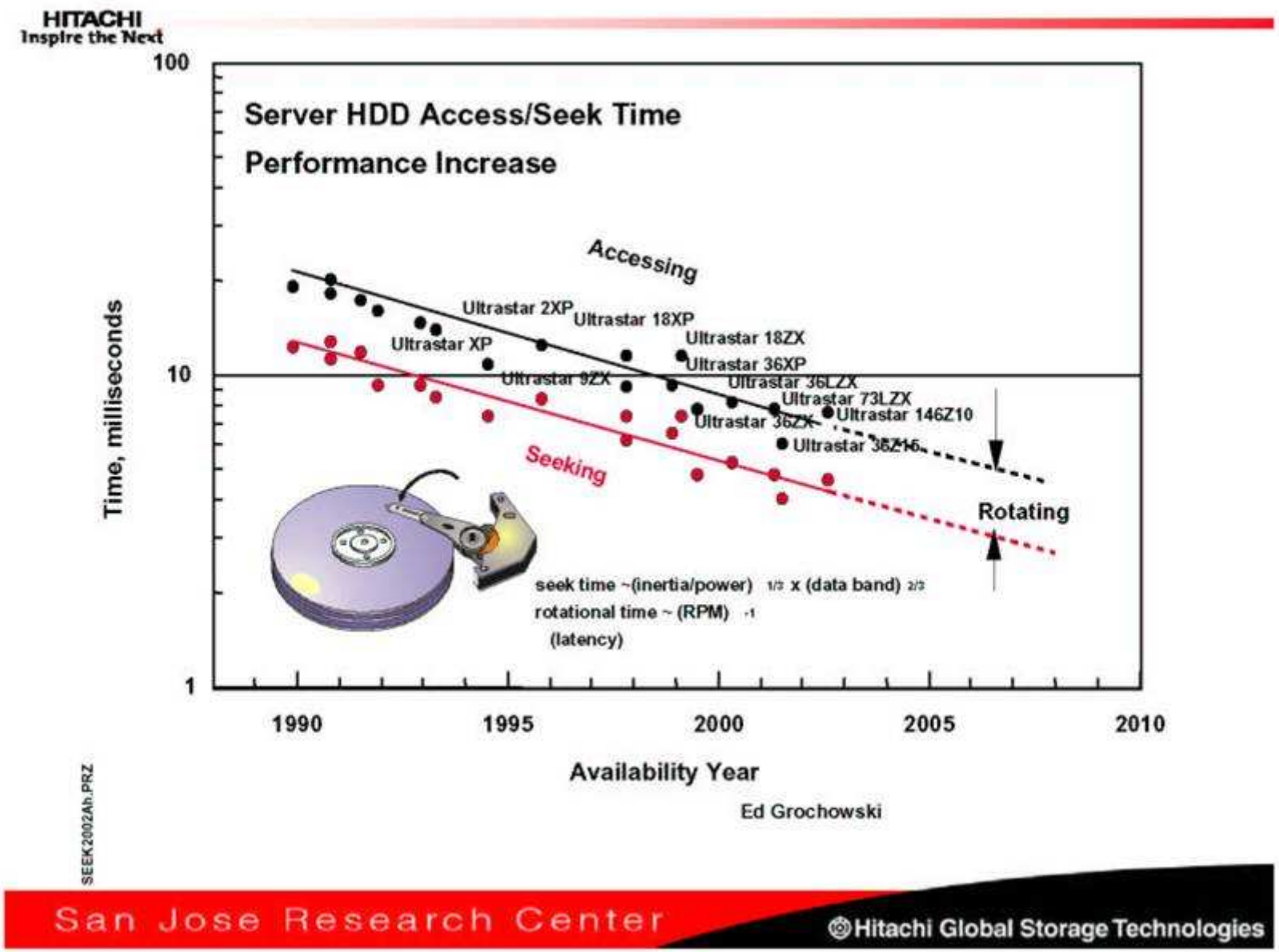
Hitachi Global Storage Technologies

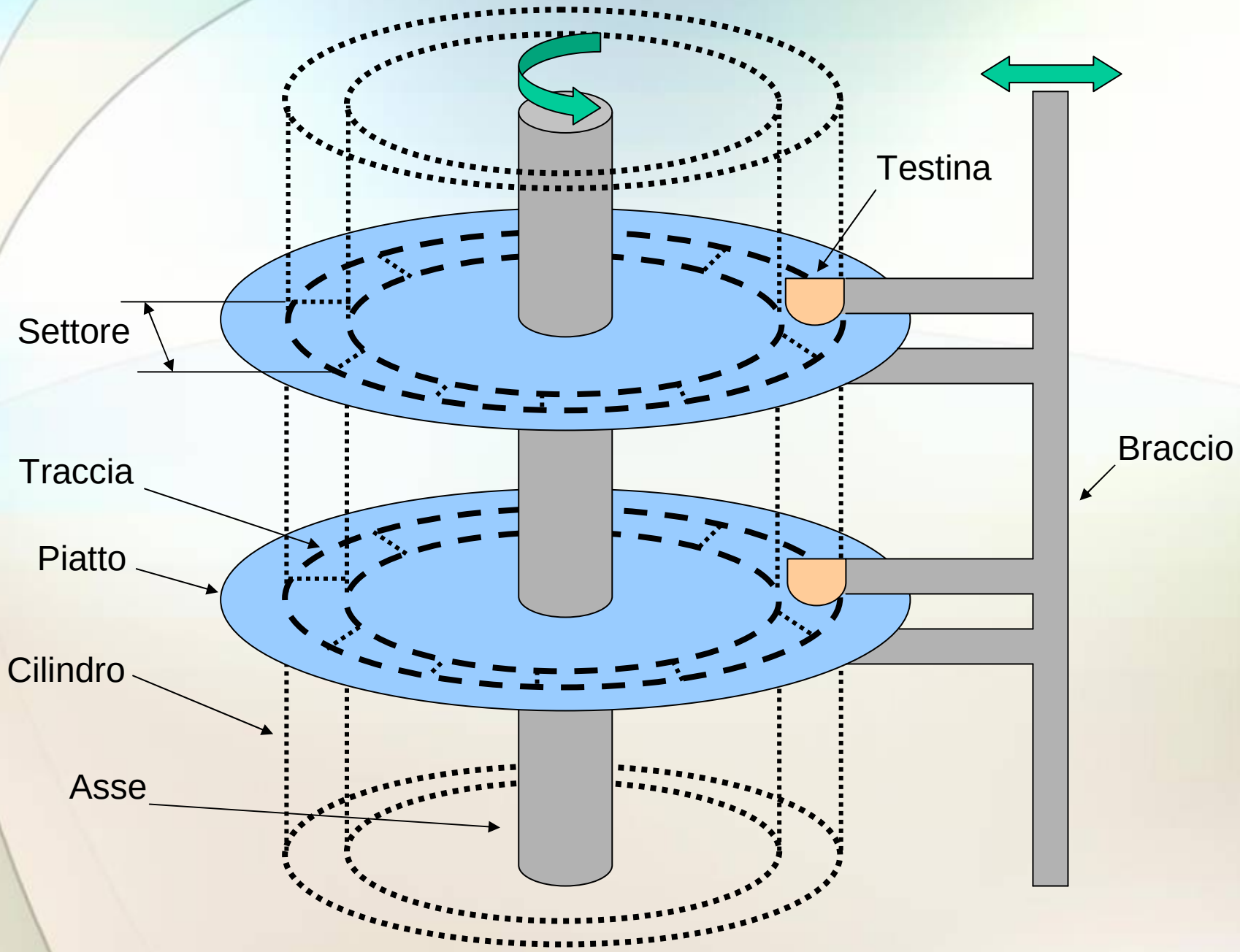
HITACHI
Inspire the Next



San Jose Research Center

Hitachi Global Storage Technologies


HITACHI
Inspire the Next
Server HDD Access/Seek Time
Performance Increase
Accessing
Seeking
Rotating
Ultrastar 2XP
Ultrastar 18XP
Ultrastar XP
Ultrastar 18ZX
Ultrastar 36XP
Ultrastar 9ZX
Ultrastar 36LZX
Ultrastar 73LZX
Ultrastar 146Z10
Ultrastar 36ZX
Ultrastar 36Z15
seek time ~(inertia/power) 1/3 x (data band) 2/3
rotational time ~ (RPM) -1
(latency)
Time, milliseconds
100
10
1
1990
1995
2000
2005
2010
Availability Year
Ed Grochowski
SEEK2002Ah.PRZ
San Jose Research Center
Hitachi Global Storage Technologies


Testina
Settore
Braccio
Traccia
Piatto
Cilindro
Asse

Elementi caratteristici di un disco

- Piatti / Testine
- Cilindri / Tracce
- Velocita' di rotazione
- Settori (numero e grandezza)
- Capacita'
- Cache
- Tipo di Bus
- Seek Time, Latency Time, Transfer Rate

Tipologie di Storage

- Ovvero come e' reso disponibile lo storage
  - DAS: Direct Attached Storage
    - I device fisici (dischi/nastri) sono collegati direttamente al server che li utilizza per l'applicazione finale
  - SAN: Storage Area Network
    - I device fisici sono collegati ad un sistema dedicato che serve dei dispositivi logici (LUN; risultato di aggregazione o partizione dei device fisici) ai server che li utilizzano
    - Generalmente le SAN forniscono anche servizi per il Cloning, Mirroring, Snapshot, Backup/Restore, Disaster Recovery per le LUN o per interi RAID group
  - NAS: Network Attached Storage
    - I device fisici sono collegati ad un sistema dedicato che serve lo spazio di storage tramite protocolli di rete (ad es.: NFS, SMB, CIFS, NCP, HTTP)

Bus

- IDE, EIDE, ATA, SATA
  - **EIDE: Enhanced Integrated Drive Electronics**

    SATA: Serial Advanced Technology Attachment
- SCSI – SAS - iSCSI
  - **SCSI: Small Computer System Interface**

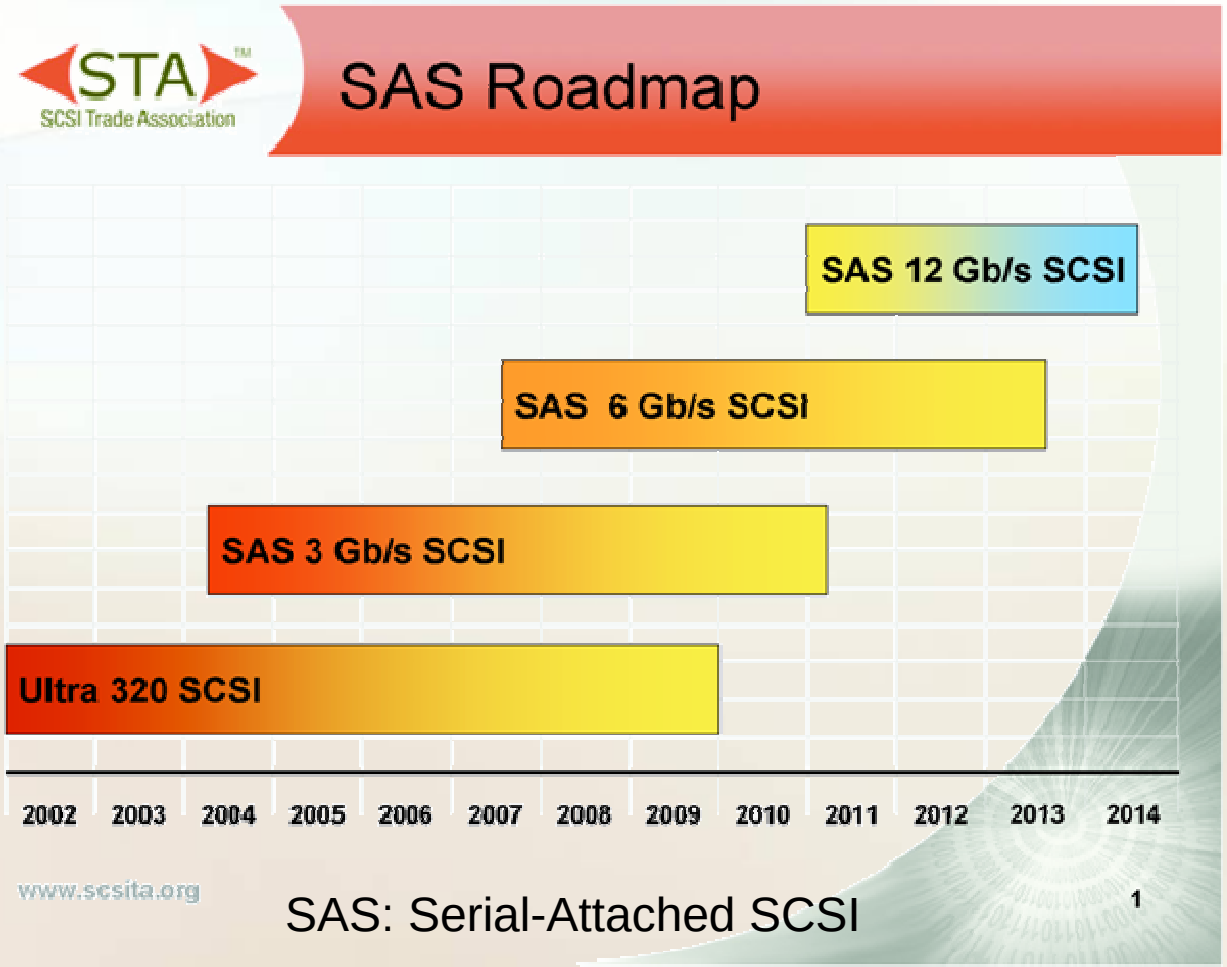
    http://www.scsita.org/terms/scsiterms.html
- SSA
  - **Serial Storage Architecture**
- Fiber Channel (copper, optical)
  - **http://www.fibrechannel.org/OVERVIEW/Roadmap.html**
  - **Collegamenti fino a qualche km su fibra ottica!**

Bus – Evoluzione SCSI (da SCSI-1 a Ultra320)

Up to 5 Megabytes Per Second		SCSI-1		Narrow (8 bit data) bus only
Up to 10 Megabytes Per Second		SCSI-2 SE	SCSI-2 Fast Differertial HVD SCSI	
Up to 20 Megabytes Per Second		SCSI-3 SPI Fast & Wide SE	SCSI-3 SPI Fast Wide Differertial HVD SCSI	Wide (16 bit data) and narrow (8 bit data) bus
Up to 40 Megabytes Per Second		Fast-20 Ultra SCSI SE	Fast-20 Ultra SCSI Differential HVD	
Up to 80 Megabytes Per Second	Ultra2 SCSI Fast-40 SPI-2 LVD SCSI			
Up to 160 Megabytes Per Second	Ultra3 or Ultra160 SCSI Fast-80 SPI-3 LVD SCSI			Wide (16 bit data) bus only
Up to 320 Megabytes Per Second	Ultra320 SCSI Fast-160 SPI-4 LVD SCSI			

http://www.3dscc.com/scsi.htm

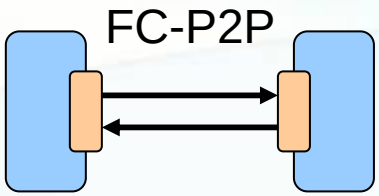
Bus – Da SCSI Parallelo a Seriale

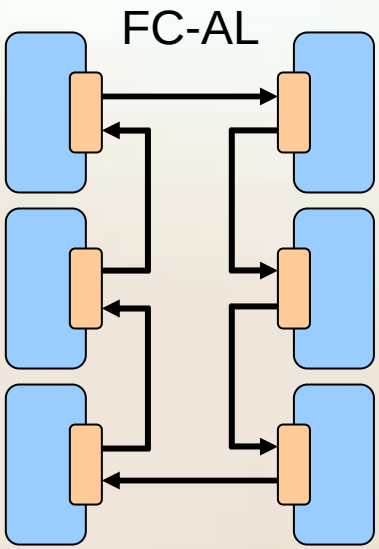


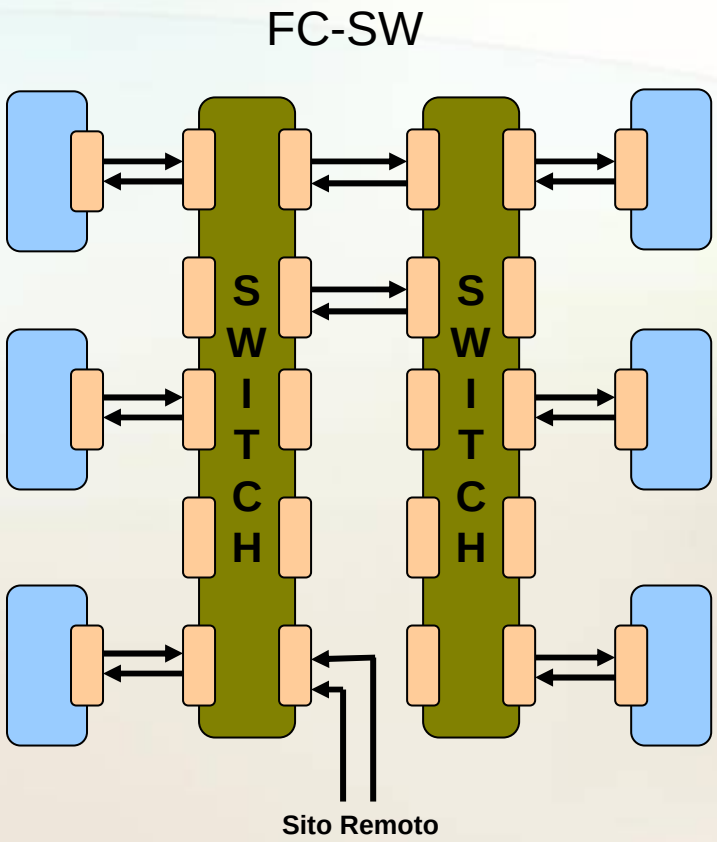
Bus – Fiber Channel - Topologie

- Point-to-Point (FC-P2P)
  - **Due dispositivi collegati direttamente**
- Arbitrated Loop (FC-AL)
  - **Fino a 127 dispositivi collegati in un "loop" (simile ad una rete di tipo "token ring").**
    Ogni device e' individuato da un indirizzo (AL-PA).
    Un algoritmo gestisce la comunicazione tra i dispositivi.
- Switched Fabric (FC-SW)
  - **Tutti i dispositivi (fino a 2^24) sono connessi a switch fiber channel (simili a switch di rete con vlan) e sono individuati da un indirizzo fisico simile ai MAC-address delle schede di rete (WWNN: World Wide Node Name). A differenza delle topologie precedenti piu' coppie di dispositivi possono scambiare dati contemporaneamente.**

Bus – Fiber Channel - Topologie



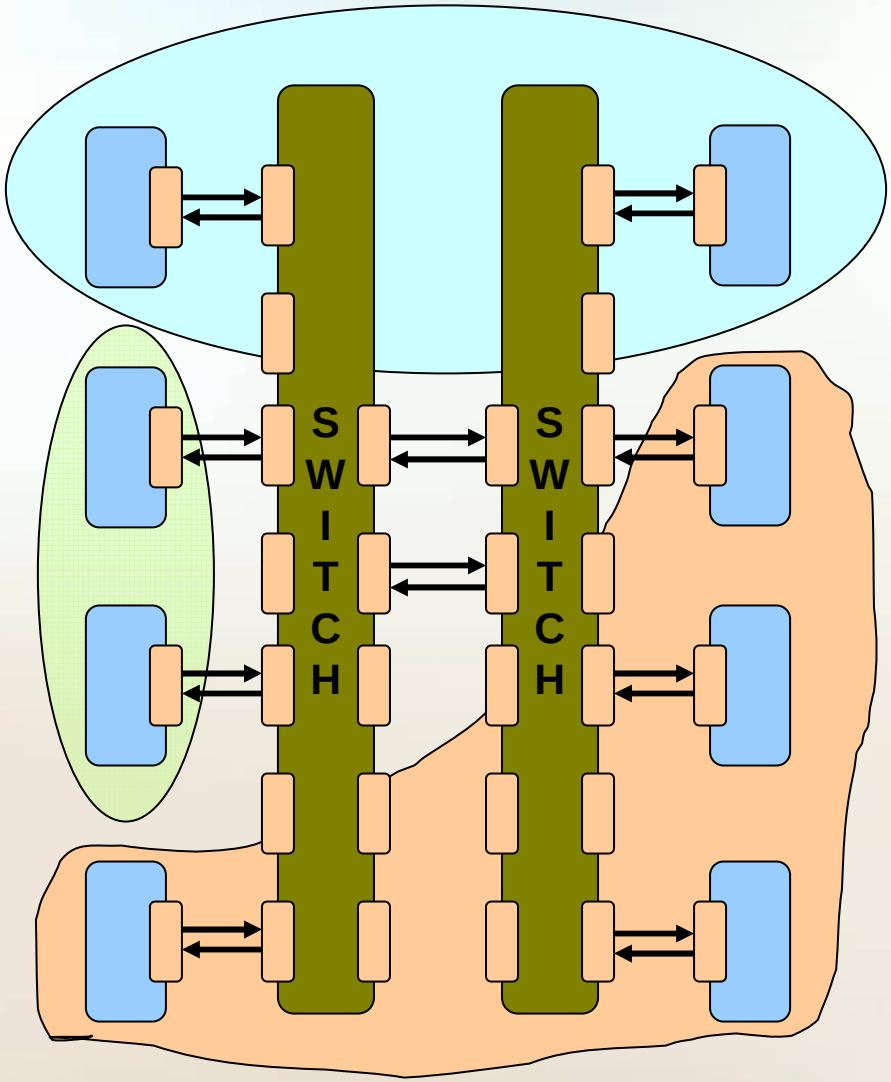




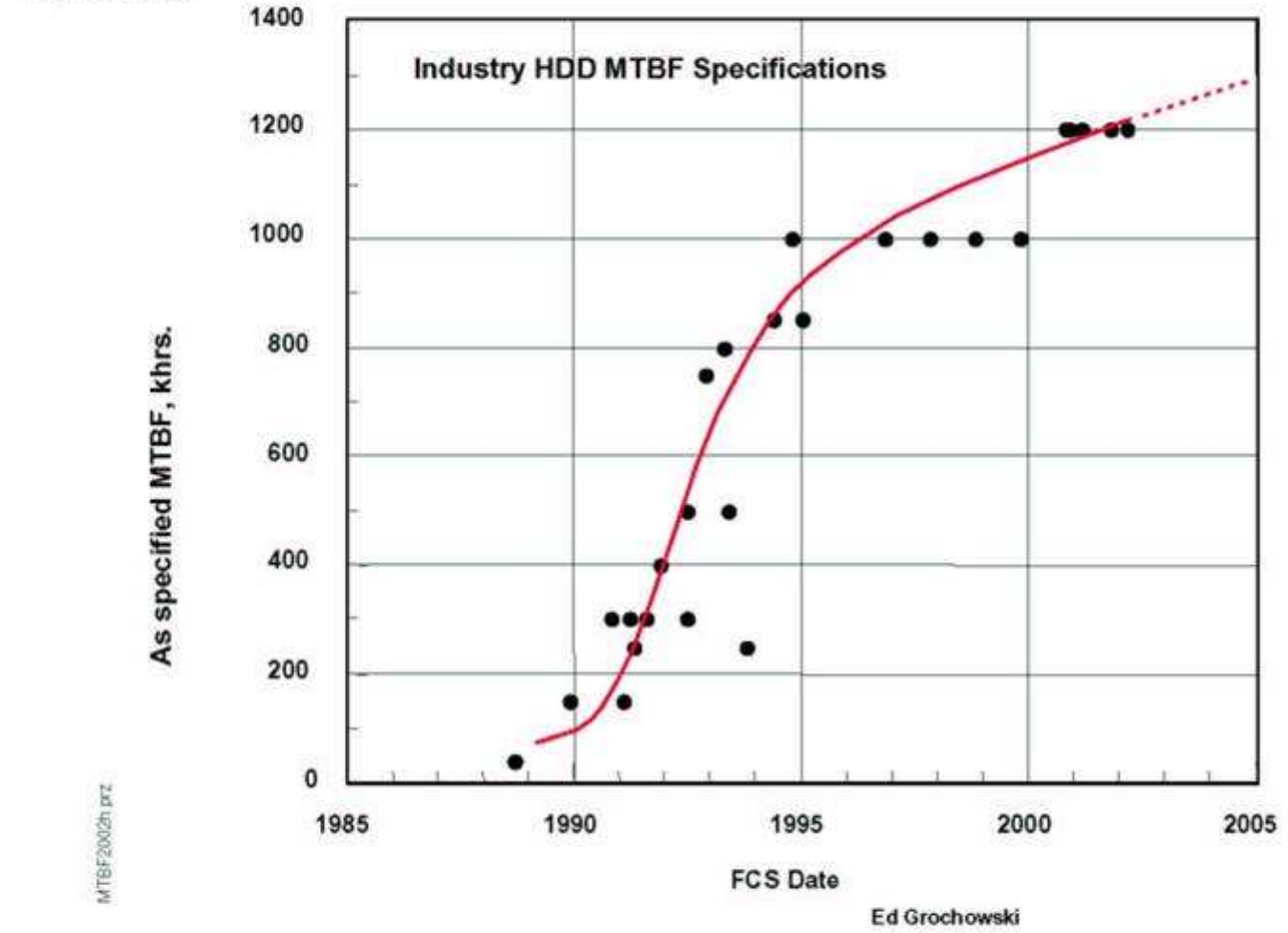
Bus – Fiber Channel - Zoning

- Nella topologia FC-SW e' possibile utilizzare lo "zoning" per limitare la visibilita' delle risorse a sottoinsiemi di dispositivi. Questo permette di semplificare la gestione e implementare politiche di sicurezza per evitare accessi da parte di dispositivi non autorizzati.
  - Soft Zoning:
    - Ogni dispositivo "vede" solo i servizi disponibili ma puo' continuare a stabilire connessioni con tutti
  - Hard Zoning:
    - Lo switch impedisce ogni comunicazione tra dispositivi appartenenti a zone diverse
  - Port Zoning:
    - Ogni porta dello switch e' associata ad una zona, qualunque dispositivo su tale porta appartera' alla zona assegnata alla porta
  - Name Zoning:
    - L'assegnazione alla zona avviene tramite l'indirizzo WWNN del dispositivo

Bus – Fiber Channel - Zoning



HITACHI
Inspire the Next



San Jose Research Center

Hitachi Global Storage Technologies

Disk Failure…

”Un disco ha solo due possibili stati:

o è rotto

o sta per rompersi !”

-- Andrei Maslennikov, Caspur

Dal JBOD al RAIDn

- JBOD: Just a Bunch Of Disks/Drives

- RAID: Redundant Array of Independent (Inexpensive) Disks
  - Aggregazione di due o piu' dischi ai fini di realizzare dei sistemi "fault tolerance" e/o aumentare le prestazioni di I/O
  - Il primo brevetto e' del 1978 (IBM)
  - Puo' essere realizzato in SW o in HW

Livelli RAID

- RAID0 (Striping)
  - Nessuna ridondanza, migliora le prestazioni di I/O ma la perdita di un solo disco determina la perdita di tutti i dati
  - MTBF inversamente proporzionale al numero di dischi!
  - Necessari almeno 2 dischi
- RAID1 (Mirroring)
  - I dati vengono scritti due (o piu' volte) su dischi diversi
  - la perdita di n-1 copie dei dati non comporta perdita di informazioni
  - Necessari almeno 2 dischi
- RAID2 (Praticamente non usato)
  - Realizza su dischi il meccanismo di correzione degli errori delle RAM ECC (Hamming Code)

Livelli RAID

- RAID3 (Bit Interleaved Parity; Raramente usato)
  - I dati vengono distribuiti su n-1 dischi come per lo striping ma viene utilizzato l'n-esimo disco per mantenere la parita'
  - La parita' si trova sempre sullo stesso disco che quindi diviene il collo di bottiglia del sistema, inoltre generalmente il RAID3 non permette operazioni di I/O multiple
  - Il guasto di un disco non comporta perdita di dati
  - Necessari almeno 3 dischi
- RAID4 (Block Interleaved Parity; Raramente usato)
  - Simile al RAID3 ma lo striping e la parita' sono a livello di blocco dati migliorando le performance e consentendo operazioni di lettura in contemporanea
- RAID5
  - Si tratta di un block striping con parita' distribuita su tutti i dischi, cio' permette di migliorare le performance e di gestire richieste di I/O multiple mantenendo l'integrita' dei dati anche in caso di guasto di un disco
- RAID6
  - Block striping con doppia parita'
  - La parita' e' realizzata su dischi differenti con due diversi metodi (ad es.: XOR e Reed-Solomon ECC code)
  - Il fail di due dischi non compromette l'integrita' dei dati
  - Necessari almeno 4 dischi

Livelli RAID “Proprietari”

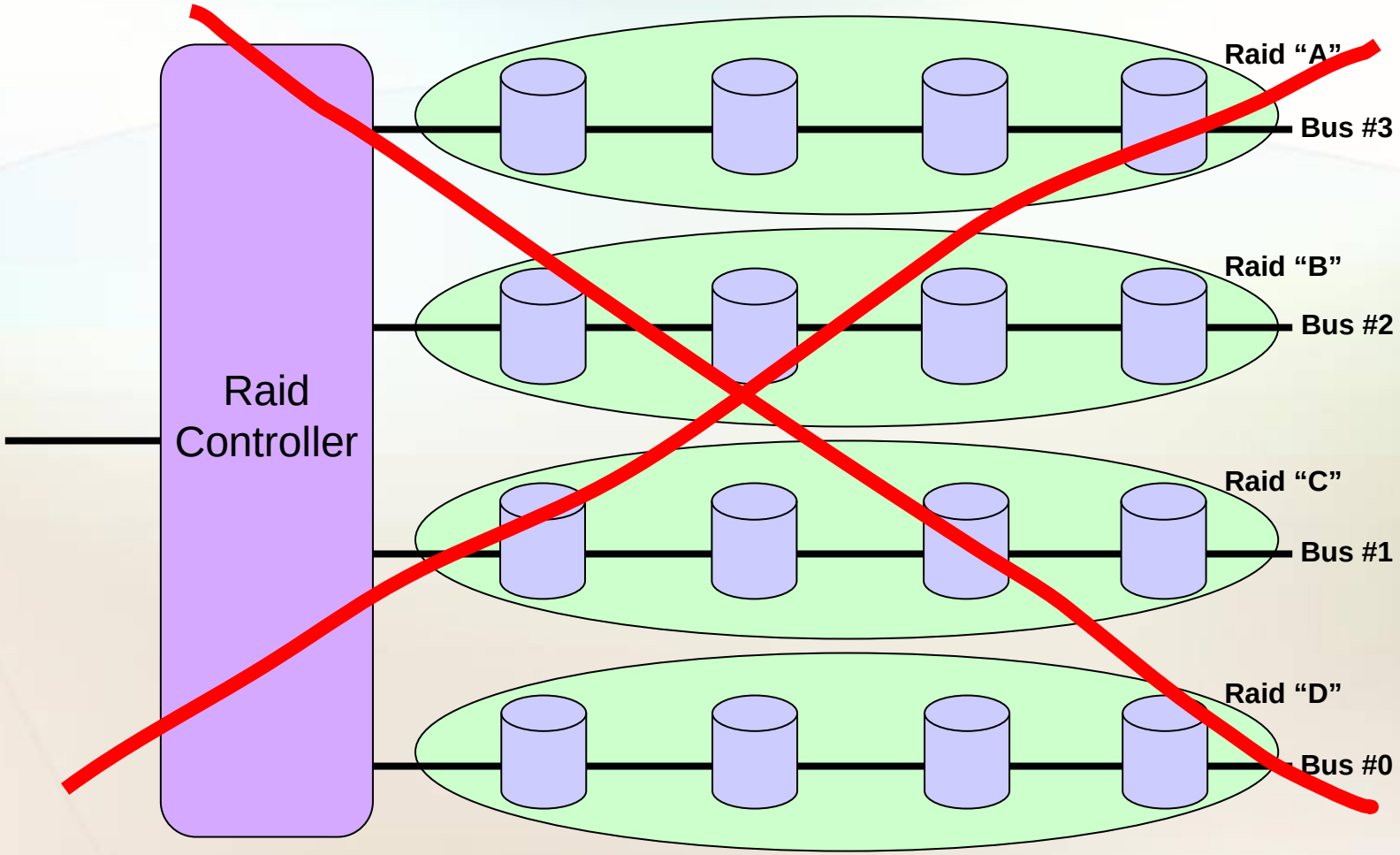
- Esistono alcune implementazioni RAID proprietarie, alcuni esempi sono:

- RAID7
  - Realizzato dalla Storage Computer Corporation.
  - Uitilizza RAID3 o RAID4 aggiungendo un sistema di caching (in RAM) per migliorare le performance
- RAID S (o Parity RAID)
  - Realizzato dalla EMC^2
  - Variante del RAID4
- Matrix RAID
  - Su due dischi utilizza una parte per realizzare un RAID0 e un’altra per realizzare un RAID1
- RAID 1.5
  - Realizzato dalla HighPoint e’ una implementazione modificata del RAID1
- Double Parity
  - Simile al RAID6 ma le parita’ sono calcolate su gruppi di blocchi diversi

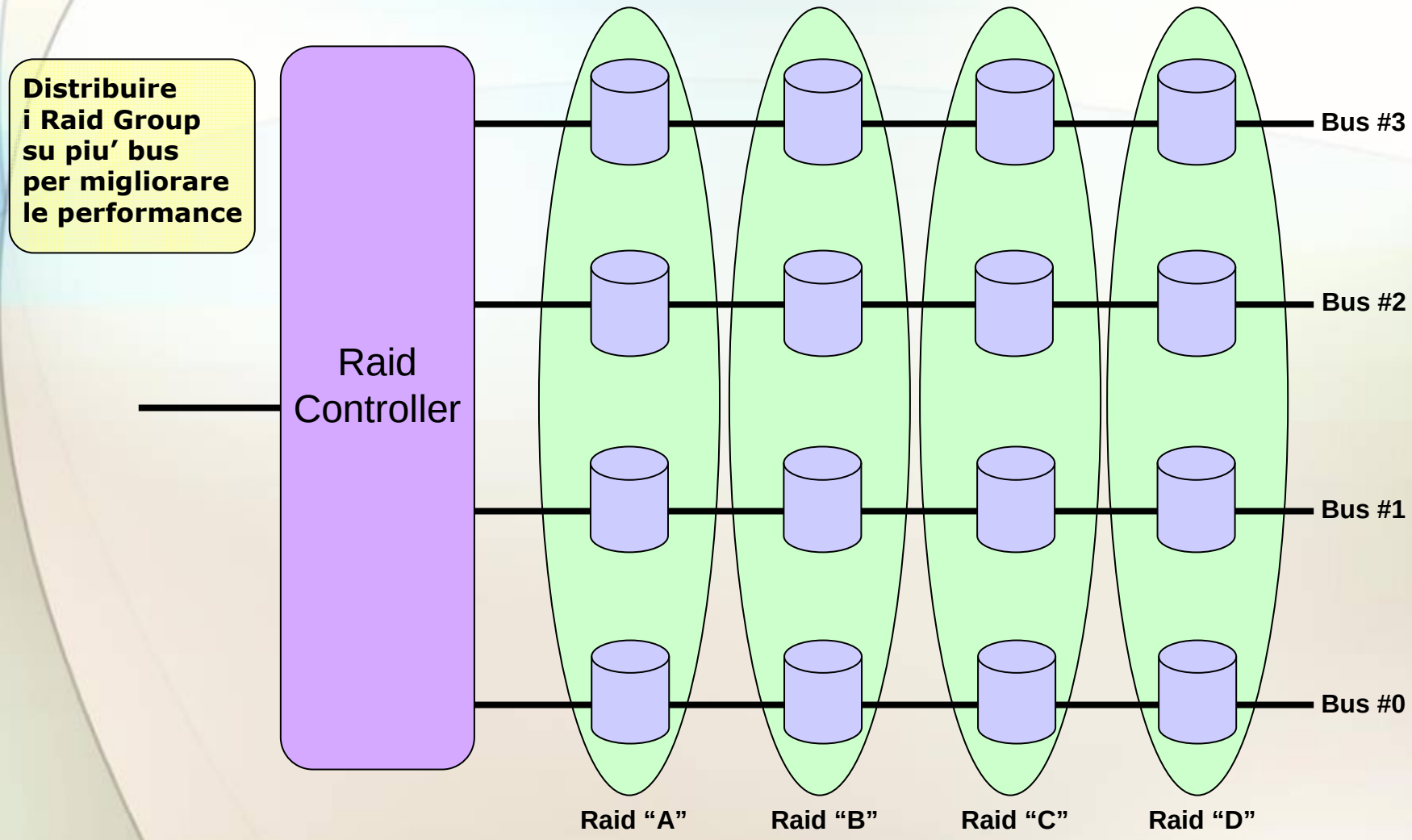
Livelli RAID “composti”

- RAID01 (RAID 0+1)
  - Mirror (RAID1) di Stripes (RAID0)
  - Necessari almeno 4 dischi
- RAID10 (RAID 1+0)
  - Stripes (RAID0) di Mirror (RAID1)
  - Necessari almeno 4 dischi
- RAID50 (RAID 5+0)
  - Stripes (RAID0) di RAID5
  - Necessari almeno 6 dischi

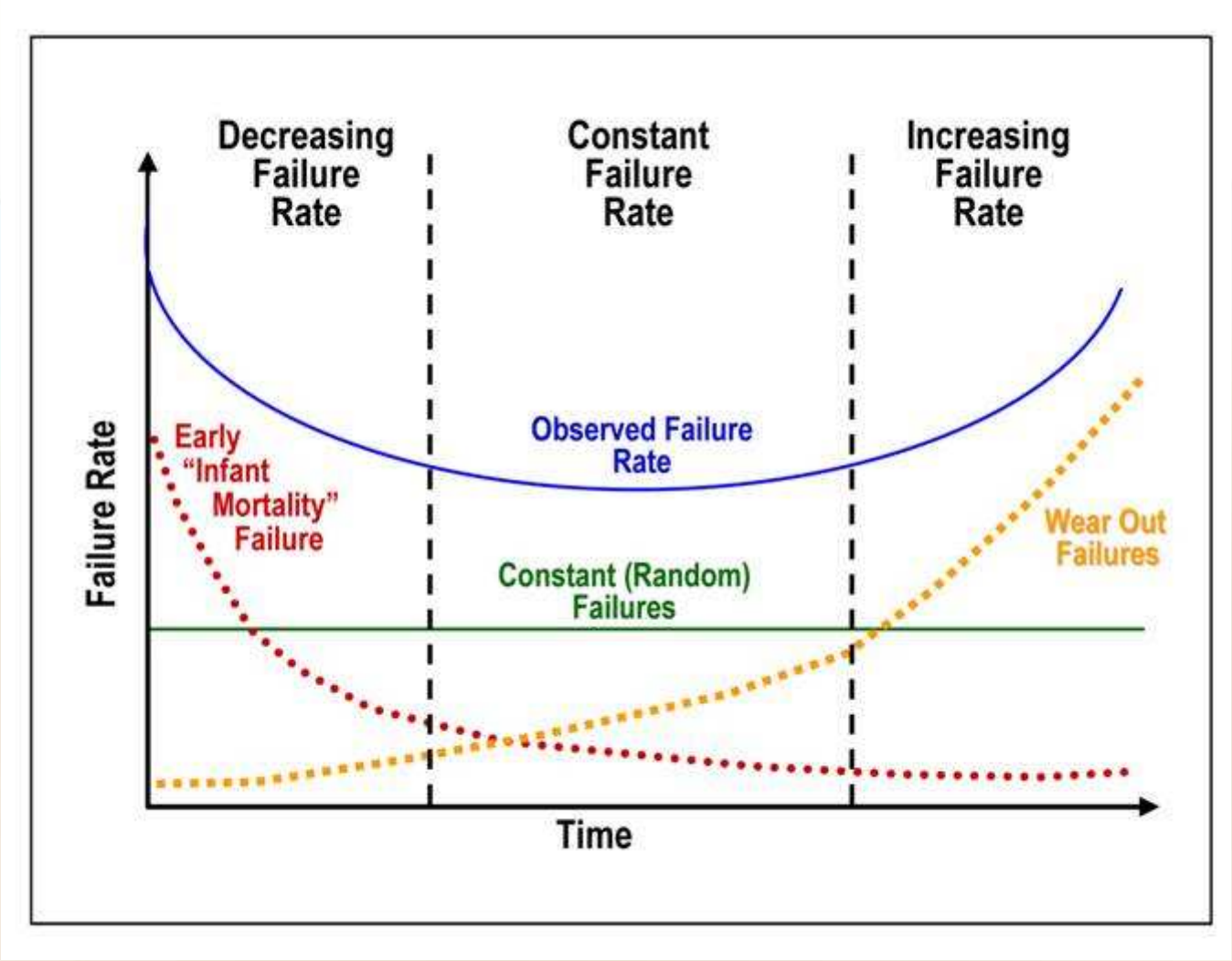
Organizzazione RAID su piu' Bus



Organizzazione RAID su piu' Bus



Bathtub Curve



Disk Failure…

- In ogni caso quando si verifica un disk failure occorre minimizzare il tempo di esposizione del sistema RAID (Stato “Degraded”)
- Il rebuilding e’ impegnativo:
  - Tempo necessario a ripristinare la situazione di fault tolerance
  - Impatto sulle performance del sistema RAID
  - Stress meccanico dei dischi “superstiti”
  - Generalmente e’ possibile modificare la priorita’ di rebuilding
- Sostituzione del disco

Disco di “scorta”: Dove e Come?

- Sistemi Hot-Swap
  - Essenziali per poter sostituire un disco senza dover spegnere l’intero sistema e fermare i servizi ad esso connessi
- Hot Spare Disk
  - Il disco e’ allocato e pronto all’uso
  - Il sistema lo utilizzera’ per la ricostruzione al posto del disco guasto
- Warm Spare Disk
  - Simile all’Hot Spare ma prima di poter utilizzare il disco il sistema RAID deve attendere lo spin-up del disco
- Global (Hot/Warm) Spare Disk
  - Il disco e’ utilizzabile in sostituzione di un disco guasto indipendentemente del “bay” di appartenenza
- “Cold” Spare Disk
  - “Classico” disco di scorta nel cassetto: quando serve non si trova…
- No Spare Disk
  - Memorandum: Un RAID5 con un disco guasto e’ un RAID0…

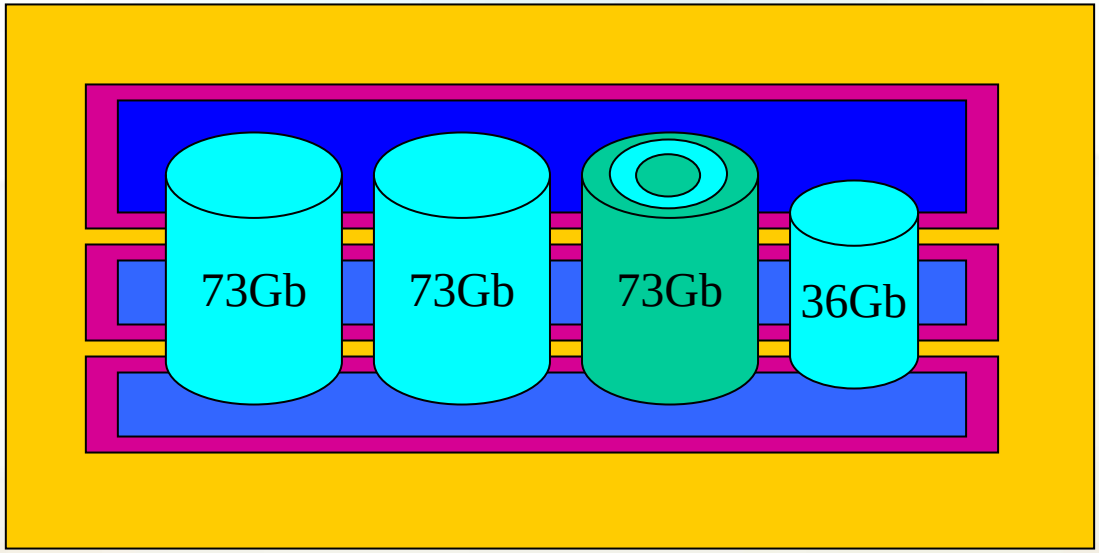
Non solo disk failure…

- Data Availability
  - I livelli RAID (diversi dal RAID0) permettono di migliorare la salvaguardia dei dati in caso di guasto di uno o piu' dischi ma nulla possono fare se il controller RAID e' unico e si guasta
  - Per migliorare la continuita' del servizio un primo passo consiste nell'utilizzare dei controller RAID doppi, generalmente disponibili da qualsiasi fabbricante di storage in due modalita':
    - Failover Active-Passive
      - Il secondo controller e' in stato di "idle", diviene attivo in seguito al fault del primo
    - Failover Active-Active
      - Entrambi i controller sono attivi e capaci di servire l'intero storage
      - Questa soluzione, piu' costosa, permette di migliorare le performance di I/O dell'intero sistema

LVM: Logical Volume Manager

- E’ una applicazione kernel-based che permette l’aggregazione logica di dispositivi fisici quali Partizioni di Hard Disk, RAID, SAN LUNs in Volumi
- Permette di migliorare le operazioni di management in quanto rende possibile aggiungere o rimuovere dinamicamente device fisici senza interrompere la disponibilita’ dello storage
  - E’ piu’ semplice, ad esempio, sostituire i dischi con altri di maggiore capacita’ senza interruzioni di servizio

LVM: Logical Volume Manager

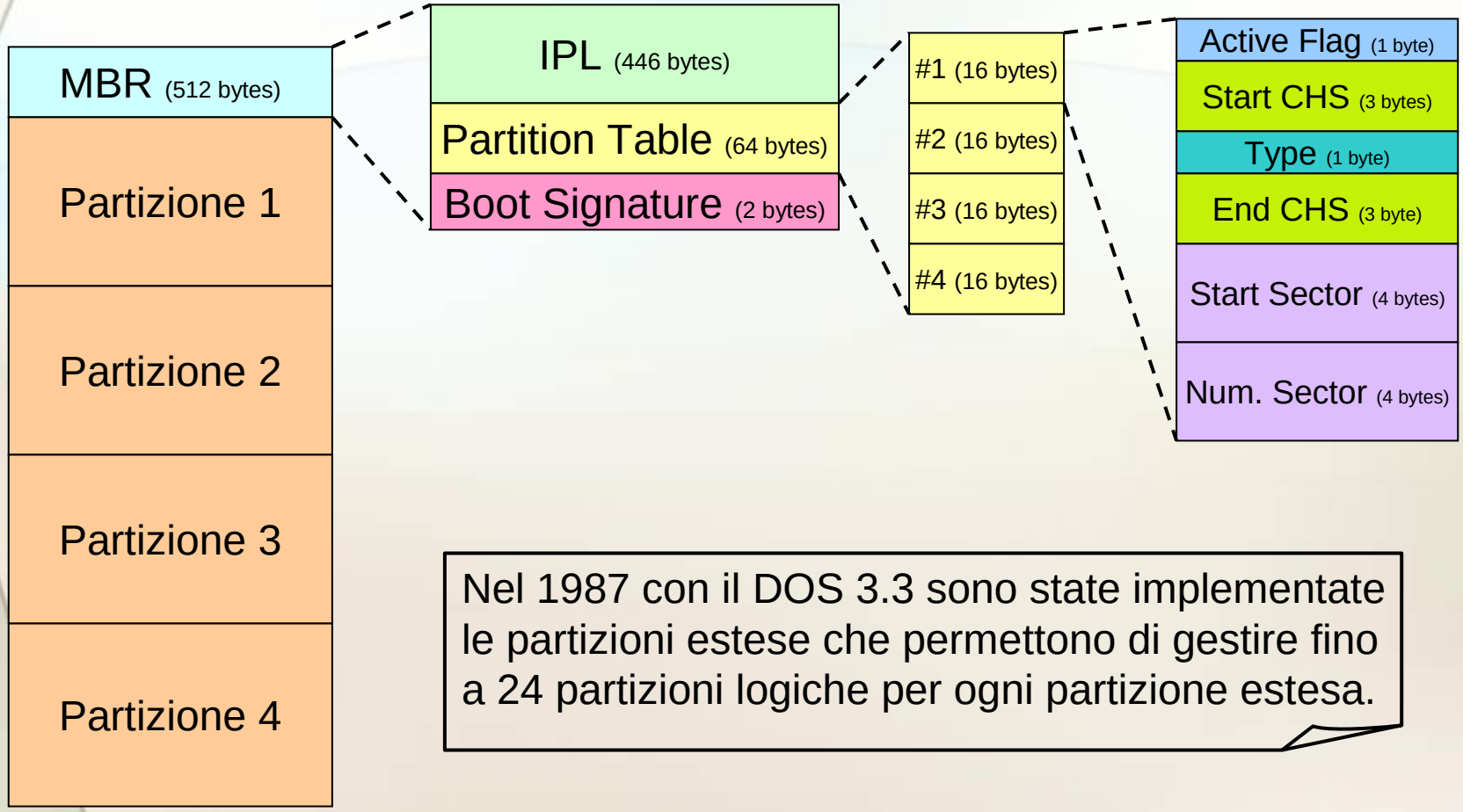


- Physical Volume: Disco fisico o un RAID array (pvcreate)
- Physical Extent: Partizioni di disco (anche un intero disco, fdisk)
- Volume Group: Insieme di Physical Volumes (vgcreate)
- Logical Volume: Insieme di Physical Extent (lvcreate)
- File System (mkfs)

Partizionamento dei dischi

- Permette di avere piu' file system sullo stesso dispositivo (JBOD o RAID)
- Esistono varie implementazioni per permettere il partizionamento dei dischi
- Linux eredita il sistema nato per l'architettura "IBM PC"
- Esistono diverse utility per la gestione delle partizioni, ad es: fdisk, GNU Parted

Partition Table (Architettura IBM PC del 1982)



Nel 1987 con il DOS 3.3 sono state implementate le partizioni estese che permettono di gestire fino a 24 partizioni logiche per ogni partizione estesa.

Partition Table - fdisk

```
[root@lx1:~]# fdisk -l /dev/cciss/c0d0

Disk /dev/cciss/c0d0: 18.2 GB, 18203197440 bytes
255 heads, 32 sectors/track, 4357 cylinders
Units = cylinders of 8160 * 512 = 4177920 bytes

           Device Boot Start     End      Blocks   Id   System
/dev/cciss/c0d0p1   *       1    2573    10497824   83   Linux
/dev/cciss/c0d0p2        2574    3063     1999200   82   Linux swap
/dev/cciss/c0d0p3        3064    4043     3998400   83   Linux
/dev/cciss/c0d0p4        4044    4357     1281120   83   Linux


[root@afs9 root]# fdisk -l /dev/sdg
Disk /dev/sdg: 549.7 GB, 549755813888 bytes
255 heads, 63 sectors/track, 66837 cylinders
Units = cylinders of 16065 * 512 = 8225280 bytes

   Device Boot Start     End      Blocks   Id   System
/dev/sdg1          1   66837   536868171   83   Linux
```

Partition Table (un esempio su HP TrueUnix)

```
dx# disklabel -r /dev/rrz9a

# /dev/rrz9a:
type: SCSI
disk: RZ29B
label:
flags:
bytes/sector: 512
sectors/track: 113
tracks/cylinder: 20
sectors/cylinder: 2260
cylinders: 3708
sectors/unit: 8380080
rpm: 7200
interleave: 1
trackskew: 9
cylinderskew: 16
headswitch: 0           # milliseconds
track-to-track seek: 0  # milliseconds
drivedata: 0

8 partitions:
#          size     offset    fstype   [fsize bsize   cpg]      # NOTE: values not exact
  a:     131072          0    unused        0     0             # (Cyl.    0 - 57*)
  b:     401408     131072    unused        0     0             # (Cyl.   57*- 235*)
  c:    8380080          0    4.2BSD     1024  8192    16       # (Cyl.    0 - 3707)
  d:    2623488     532480    unused        0     0             # (Cyl.  235*- 1396*)
  e:    2623488    3155968    unused        0     0             # (Cyl. 1396*- 2557*)
  f:    2600624    5779456    unused        0     0             # (Cyl. 2557*- 3707)
  g:    3936256     532480    unused        0     0             # (Cyl.  235*- 1977*)
  h:    3911344    4468736    unused        0     0             # (Cyl. 1977*- 3707)
```

Utility per il controllo dei dischi

- Ci possono essere diverse possibilita' per controllaro la "bonta'" di un disco:
  - Utility incluse nel BIOS
    - Utilizzabili solo a sistema fermo
  - Utility Stand-Alone
    - Di solito un floppy o un CD
    - Utilizzabili solo a sistema fermo
    - Es.: DFT (Disk Fitness Test) http://www.hgst.com/hdd/technolo/dft/dftnew.htm
  - Utility di sistema
    - Possono richiedere l'umount del file system
    - Es.: badblocks
  - Utility del controller RAID
    - Possono richiedere l'umount del file system

Il comando badblocks

- Permette di fare lo scan di un device per verificare la presenza di blocchi difettosi
  - Puo' essere usato in varie modalita':
    - Read-only (default)
      - I blocchi dati sono solo letti)
    - Read-Write non distruttivo
      - Il contenuto del blocco e' salvato e poi ripristinato dopo il test di scrittura e lettura con vari pattern
    - Read-Write distruttivo

“Path” di un disco (scsi, fc)

- Un server puo’ gestire una grande quantita’ di dischi, ognuno dei quali e’ individuato tramite un insieme di informazioni composte da:
  - Host
    - Identifica l’adapter (es.: scsi0)
  - Channel
    - Identifica il canale sul singolo adapter (generalmente 0 o 1)
  - Id
    - Identificativo del disco
  - LUN (Logical Unit Number)
    - Identificativo dell’unita’ logica (generalmente usato nei sistemi collegati in Storage Area Network)

Aggiunta/Rimozione Dischi

- Ormai la quasi totalita' dei sistemi operativi permettono l'aggiunta o la rimozione dei dischi "on-line"
- Abbiamo visto che con Linux il file "virtuale" /proc/scsi/scsi contiene l'elenco dei device fisici connessi al sistema
- E' possibile aggiungere o rimuovere un device, interagendo con tale file:
  - echo "scsi add-single-device HOST CHANNEL ID LUN" >> /proc/scsi/scsi
  - echo "scsi remove-single-device HOST CHANNEL ID LUN" >> proc/scsi/scsi
  - Il successo dell'operazione puo' essere controllato verificando il file /proc/scsi/scsi

Cosa “vede” il Sistema al Boot

- Interfacce IDE, SCSI, Fiber Channel, etc.
- Scan dei Logical & Physical Devices
- “Attach” dei Devices al Sistema
- Check “Partition Table” dei Dischi
- Scan per Volume Groups, Start LVM
- Scan per MD (Multiple Disk)

Esempio di un sistema DAS (Direct Attached Storage)

- Dell PowerEdge 2550 (2 x P3 1GHz/256k)
- Ram 512MB 133 MHz
- RAID Perc3/Di 2 canali 128MB cache
- RAID Perc3/DC 2 canali 128MB cache
- 2 HD 18GB SCSI Ultra 3 10 Krpm
- 3 HD 73GB SCSI Ultra 3 10 Krpm
- Dell PV220S + 14HD 73GB SCSI Ultra3 10Krpm

```
afs4 kernel: megaraid: [161J:3.17] detected 10 logical drives
afs4 kernel: megaraid: channel[1] is raid.
afs4 kernel: megaraid: channel[2] is raid.
afs4 kernel: scsi0 : LSI Logic MegaRAID 161J 254 commands 15 targs 5 chans 7 luns
afs4 kernel: blk: queue c2546a14, I/O limit 4095Mb (mask 0xffffffff)
afs4 kernel: scsi0: scanning virtual channel 0 for logical drives.
afs4 kernel:   Vendor: MegaRAID  Model: LD0 RAID0 69878R  Rev: 161J
afs4 kernel:   Type:   Direct-Access                      ANSI SCSI revision: 02
afs4 kernel: blk: queue c2546c14, I/O limit 4095Mb (mask 0xffffffff)
afs4 kernel:   Vendor: MegaRAID  Model: LD1 RAID0 69878R  Rev: 161J
afs4 kernel:   Type:   Direct-Access                      ANSI SCSI revision: 02
afs4 kernel: blk: queue dfc00014, I/O limit 4095Mb (mask 0xffffffff)

         [… omissis …]

afs4 kernel:   Vendor: MegaRAID  Model: LD7 RAID0 39756R  Rev: 161J
afs4 kernel:   Type:   Direct-Access                      ANSI SCSI revision: 02
afs4 kernel: blk: queue dfb9c814, I/O limit 4095Mb (mask 0xffffffff)
afs4 kernel:   Vendor: MegaRAID  Model: LD8 RAID1 69878R  Rev: 161J
afs4 kernel:   Type:   Direct-Access                      ANSI SCSI revision: 02
afs4 kernel: blk: queue dfb9cc14, I/O limit 4095Mb (mask 0xffffffff)
afs4 kernel:   Vendor: MegaRAID  Model: LD9 RAID5 39756R  Rev: 161J
afs4 kernel:   Type:   Direct-Access                      ANSI SCSI revision: 02
afs4 kernel: blk: queue dfb38014, I/O limit 4095Mb (mask 0xffffffff)
afs4 kernel: scsi0: scanning virtual channel 1 for logical drives.
afs4 kernel: scsi0: scanning virtual channel 2 for logical drives.
afs4 kernel: scsi0: scanning physical channel 0 for devices.
afs4 kernel:   Vendor: DELL      Model: PV22XS            Rev: E.10
afs4 kernel:   Type:   Processor                          ANSI SCSI revision: 03
afs4 kernel: blk: queue dfb18414, I/O limit 4095Mb (mask 0xffffffff)
afs4 kernel: scsi0: scanning physical channel 1 for devices.
afs4 kernel:   Vendor: DELL      Model: PV22XS            Rev: E.10
afs4 kernel:   Type:   Processor                          ANSI SCSI revision: 03
afs4 kernel: blk: queue dfafd214, I/O limit 4095Mb (mask 0xffffffff)
```

```
afs4 kernel: Attached scsi disk sda at scsi0, channel 0, id 0, lun 0
afs4 kernel: Attached scsi disk sdb at scsi0, channel 0, id 1, lun 0
afs4 kernel: Attached scsi disk sdc at scsi0, channel 0, id 2, lun 0
afs4 kernel: Attached scsi disk sdd at scsi0, channel 0, id 3, lun 0
afs4 kernel: Attached scsi disk sde at scsi0, channel 0, id 4, lun 0
afs4 kernel: Attached scsi disk sdf at scsi0, channel 0, id 5, lun 0
afs4 kernel: Attached scsi disk sdg at scsi0, channel 0, id 6, lun 0
afs4 kernel: Attached scsi disk sdh at scsi0, channel 0, id 8, lun 0
afs4 kernel: Attached scsi disk sdi at scsi0, channel 0, id 9, lun 0
afs4 kernel: Attached scsi disk sdj at scsi0, channel 0, id 10, lun 0

          [… omissis …]

afs4 kernel: SCSI device sda: 143110144 512-byte hdwr sectors (73272 MB)
afs4 kernel: Partition check:
afs4 kernel:  sda: unknown partition table

          [… omissis …]

afs4 kernel: SCSI device sdb: 143110144 512-byte hdwr sectors (73272 MB)
afs4 kernel: SCSI device sdc: 143110144 512-byte hdwr sectors (73272 MB)
afs4 kernel: SCSI device sdd: 143110144 512-byte hdwr sectors (73272 MB)
afs4 kernel: SCSI device sde: 143110144 512-byte hdwr sectors (73272 MB)
afs4 kernel: SCSI device sdf: 143110144 512-byte hdwr sectors (73272 MB)
afs4 kernel: SCSI device sdg: 143110144 512-byte hdwr sectors (73272 MB)
afs4 kernel: SCSI device sdh: 286220288 512-byte hdwr sectors (146545 MB)
afs4 kernel: SCSI device sdi: 143110144 512-byte hdwr sectors (73272 MB)
afs4 kernel: SCSI device sdj: 286220288 512-byte hdwr sectors (146545 MB)
```

```
afs4 kernel: Red Hat/Adaptec aacraid driver, Sep  4 2002

          [… omissis …]

afs4 kernel: scsi1 : percraid
afs4 kernel:   Vendor: DELL      Model: PERCRAID Mirror   Rev: V1.0
afs4 kernel:   Type:   Direct-Access                      ANSI SCSI revision: 02
afs4 kernel:   Vendor: DELL      Model: PERCRAID RAID5    Rev: V1.0
afs4 kernel:   Type:   Direct-Access                      ANSI SCSI revision: 02

          [… omissis …]

afs4 kernel: Attached scsi removable disk sdk at scsi1, channel 0, id 0, lun 0
afs4 kernel: Attached scsi removable disk sdl at scsi1, channel 0, id 2, lun 0
afs4 kernel: SCSI device sdk: 35544576 512-byte hdwr sectors (18199 MB)
afs4 kernel: sdk: Write Protect is off
afs4 kernel:  sdk: sdk1 sdk2 sdk3
afs4 kernel: SCSI device sdl: 286714368 512-byte hdwr sectors (146798 MB)
afs4 kernel: sdl: Write Protect is off
afs4 kernel:  sdl1 sdl2 sdl3 sdl4

          [… omissis …]
```

```
afs4 fsck: /: clean, 216128/749248 files, 1088464/1498061 blocks
afs4 rc.sysinit: Checking root filesystem succeeded
afs4 rc.sysinit: Remounting root filesystem in read-write mode:  succeeded

        [… omissis …]

afs4 vgscan: vgscan -- reading all physical volumes (this may take a while...)
afs4 vgscan: vgscan -- found inactive volume group mastervg
afs4 vgscan: vgscan -- /etc/lvmtab and /etc/lvmtab.d successfully created
afs4 vgscan: vgscan -- WARNING: This program does not do a VGDA backup of your volume group
afs4 rc.sysinit: Setting up Logical Volume Management: succeeded
afs4 rc.sysinit: Activating swap partitions:  succeeded
afs4 vgscan: vgscan -- found active volume group mastervg
afs4 rc.sysinit: Setting up Logical Volume Management: succeeded

        [… omissis …]


afs4 rc.sysinit: Checking filesystems succeeded
afs4 rc.sysinit: Mounting local filesystems:  succeeded
afs4 rc.sysinit: Enabling local filesystem quotas:  succeeded
afs4 rc.sysinit: Enabling swap space:  succeeded
```

E dopo il boot….(1)

- **Elenco dischi SCSI visti dal sistema:**
  - **cat /proc/scsi/scsi**

```
Attached devices:
Host: scsi0 Channel: 00 Id: 06 Lun: 00
  Vendor: DELL     Model: PV22XS           Rev: E.10
  Type:   Processor                        ANSI SCSI revision: 03
Host: scsi0 Channel: 01 Id: 06 Lun: 00
  Vendor: DELL     Model: PV22XS           Rev: E.10
  Type:   Processor                        ANSI SCSI revision: 03

            [… omissis …]

Host: scsi1 Channel: 00 Id: 00 Lun: 00
  Vendor: DELL     Model: PERCRAID Mirror  Rev: 0001
  Type:   Direct-Access                    ANSI SCSI revision: 02
Host: scsi1 Channel: 00 Id: 02 Lun: 00
  Vendor: DELL     Model: PERCRAID RAID5   Rev: 0001
  Type:   Direct-Access                    ANSI SCSI revision: 02
```

E dopo il boot….(2)

- **Elenco partizioni gestite dal sistema:**
  - **cat /proc/partitions**

```
major minor  #blocks  name

  58     0   16777216 lvma
  58     1   16777216 lvmb
  58     2   16777216 lvmc
  58     3   16777216 lvmd
   8     0   71555072 sda
   8     1   71553478 sda1
   8    16   71555072 sdb
   8    17   71553478 sdb1

[… omissis …]

   9     0   71555008 md0
```

E dopo il boot….(3)

- Con controller “MegaRAID” ulteriori info in:
  - /proc/megaraid/<Adapter>/:
    - config
    - mailbox
    - stat
    - status
  - Esempio di “cat /proc/megaraid/0/config”:
    - **PERC 3/DC**
      Controller Type: 438/466/467/471/493
      Controller Supports 40 Logical Drives
      Controller / Driver uses 64 bit memory addressing
      Base = e0837000, Irq = 20, Logical Drives = 9, Channels = 2
      Version =161J:3.17, DRAM = 128Mb
      Controller Queue Depth = 254, Driver Queue Depth = 126

E dopo il boot….(4)

- In /proc/fs/… si possono trovare informazioni su lo stato di particolari filesystem (se usati), Es.:
  - XFS: /proc/fs/xfs/stat e /proc/fs/xfs/xqm
  - ReiserFS, in /proc/fs/reiserfs/<sd(N,M)>/:
    - bitmap, per-level, journal, super, oidmap, version, on-disk-super
    - Es journal stats: cat /proc/fs/reiserfs/sd\(8\,49\)/journal

```
s_journal_block:  18
s_journal_dev:  none[0]
s_orig_journal_size:  8192
s_journal_trans_max:  1024
s_journal_block_count:  1506843122
s_journal_max_batch:  900
s_journal_max_commit_age:  30
s_journal_max_trans_age:  0
j_state:  0
j_trans_id:  4354
j_mount_id:  10
j_start:  1355
j_len:  0
…
… etc …
```

Creazione di RAID Software

- Uso dei device MD (Multiple Disk)
  - Esempi Creazione RAID sw con mdadm:
    - RAID0: mdadm --create /dev/md0 --level=0 --raid-devices=2 /dev/sda /dev/sdb
    - RAID1: mdadm --create /dev/md1 --level=1 --raid-devices=2 /dev/sdc /dev/sdd
    - RAID5: mdadm --create /dev/md2 --level=5 --raid-devices=3 /dev/sde /dev/sdf /dev/sdg
  - Verifica stato devices MD:
    - cat /proc/mdstat:

```
md2 : active raid5 sdg[3] sdf[1] sde[0]
      143110016 blocks level 5, 64k chunk, algorithm 2 [3/2] [UU_]
      [========>.............] recovery = 43.4% (31063740/71555008) finish=66.8min speed=10094K/sec
md1 : active raid1 sdd[1] sdc[0]
      71555008 blocks [2/2] [UU]
      [========>.............] resync = 43.9% (31448248/71555008) finish=64.8min speed=10313K/sec
md0 : active raid0 sdb[1] sda[0]
      143110016 blocks 64k chunks
```

```
[root@afs4 root]# mdadm --detail /dev/md0
/dev/md0:
        Version : 00.90.00
  Creation Time : Thu Jan 16 10:26:00 2003
     Raid Level : raid0
     Array Size : 143110016 (136.48 GiB 146.54 GB)
   Raid Devices : 2
  Total Devices : 2
Preferred Minor : 0
    Persistence : Superblock is persistent

    Update Time : Thu Jan 16 10:26:00 2003
          State : dirty, no-errors
 Active Devices : 2
Working Devices : 2
 Failed Devices : 0
  Spare Devices : 0

     Chunk Size : 64K

    Number   Major   Minor   RaidDevice State
       0       8        0        0      active sync   /dev/sda
       1       8       16        1      active sync   /dev/sdb
           UUID : 88775838:df173557:ae4c497c:12dba7e8
[root@afs4 root]#
```

```
[root@afs4 root]# mdadm --detail /dev/md1
/dev/md1:
        Version : 00.90.00
  Creation Time : Thu Jan 16 10:26:21 2003
     Raid Level : raid1
     Array Size : 71555008 (68.24 GiB 73.27 GB)
    Device Size : 71555008 (68.24 GiB 73.27 GB)
   Raid Devices : 2
  Total Devices : 2
Preferred Minor : 1
    Persistence : Superblock is persistent

    Update Time : Thu Jan 16 10:26:21 2003
          State : dirty, no-errors
 Active Devices : 2
Working Devices : 2
 Failed Devices : 0
  Spare Devices : 0


    Number   Major   Minor   RaidDevice State
       0       8       32        0      active sync   /dev/sdc
       1       8       48        1      active sync   /dev/sdd
           UUID : 15b638b3:3f4443de:86d03604:69060e94
[root@afs4 root]#
```

```
[root@afs4 root]# mdadm --detail /dev/md2
/dev/md2:
        Version : 00.90.00
  Creation Time : Thu Jan 16 10:26:30 2003
     Raid Level : raid5
     Array Size : 143110016 (136.48 GiB 146.54 GB)
    Device Size : 71555008 (68.24 GiB 73.27 GB)
   Raid Devices : 3
  Total Devices : 4
Preferred Minor : 2
    Persistence : Superblock is persistent

    Update Time : Thu Jan 16 11:33:19 2003
          State : dirty, no-errors
 Active Devices : 3
Working Devices : 3
 Failed Devices : 1
  Spare Devices : 0

         Layout : left-symmetric
     Chunk Size : 64K

    Number   Major   Minor   RaidDevice State
       0       8       64        0      active sync   /dev/sde
       1       8       80        1      active sync   /dev/sdf
       2       8       96        2      active sync   /dev/sdg
           UUID : bc0acf26:98f66b2a:86b1744b:45539aec
[root@afs4 root]#
```

Esempio di sistema con RAID Hardware

- Sistema Dell Poweredge 2550 con 2 RAID:
  - RAID Perc3/Di 2 canali 128MB cache **(dischi interni)**
  - RAID Perc3/DC 2 canali 128MB cache **(dischi esterni)**
- Utility fornite dal costruttore per il RAID management:

– **afaapps-2.7-2.i386.rpm**	**(afacli: gestione driver di tipo aacraid)**
– **afasnmp-2.7-1.i386.rpm**	**(agent snmp per driver aacraid)**
– **dellmgr-5.23-1.i386.rpm**	**(gestione con tool semigrafico per driver di tipo megaraid)**
– **linflash-2.14-0.i386.rpm**	**(per l'aggiornamento del firmware sulla flash ram dei controller)**
– **megamon-3.6-0.i386.rpm**	**(monitoring driver di tipo megaraid)**
– **megarc-1.07-1.i386.rpm**	**(gestione driver di tipo megaraid)**
– **percsnmp-4.06-1.i386.rpm**	**(agent snmp per controller "perc")**

RAID HW: afacli (1)

```
[root@afs4 tmp]# afacli
-----------------------------------------------------------------------
DELL PowerEdge Expandable RAID Controller 2 Command Line Interface
Copyright 1998-2002 Adaptec, Inc. All rights reserved
-----------------------------------------------------------------------
FASTCMD>
FASTCMD> controller list
Executing: controller list

Adapter Name    Adapter Type          Availability  Clustering
------------    ------------          ------------  ------------
afa0                PERC 3/Di         read/write           No

FASTCMD> open afa0
Executing: open "afa0"
AFA0>
AFA0> enclosure list
Executing: enclosure list

Enclosure
ID (B:ID:L) Fan Power Slot Sensor Door Speaker  Standard Diagnostic
----------- --- ----- ---- ------ ---- -------- -------- ----------
 0  1:06:0   0    0    3     0     0    No    SAF-TE  PASSED
 1  0:06:0   0    0    2     2     0    No    SAF-TE  PASSED

AFA0> container list
Executing: container list
Num          Total  Oth Chunk          Scsi   Partition
Label Type   Size   Ctr Size   Usage   B:ID:L Offset:Size
----- ------ ------ --- ------ ------- ------ -------------
 0    Mirror 16.9GB            Open    0:00:0 64.0KB:16.9GB
 /dev/sdj             system           0:01:0 64.0KB:16.9GB

 2    RAID-5  136GB    32KB Open    1:02:0 64.0KB:68.3GB
 /dev/sdk                               1:03:0 64.0KB:68.3GB
                                        1:04:0 64.0KB:68.3GB
```

RAID HW: afacli (2)

```
AFA0> controller details
Executing: controller details
Controller Information
----------------------
         Remote Computer: .
             Device Name: AFA0
         Controller Type: PERC 3/Di
             Access Mode: READ-WRITE
Controller Serial Number: Last Six Digits = 3010D2
         Number of Buses: 2
         Devices per Bus: 15
          Controller CPU: i960 R series
    Controller CPU Speed: 100 Mhz
       Controller Memory: 126 Mbytes
           Battery State: Ok
Component Revisions
-------------------
               CLI: 2.7-1 (Build #4944)
               API: 2.7-1 (Build #4944)
    Miniport Driver: 3.0-0 (Build #20773)
Controller Software: 2.7-0 (Build #3546)
    Controller BIOS: 2.7-0 (Build #3546)
Controller Firmware: (Build #3546)

AFA0> container show cache 2
Executing: container show cache 2

Global Container Read Cache Size  : 0
Global Container Write Cache Size : 118259712
Read Cache Setting        : ENABLE
Write Cache Setting   :     ENABLE WHEN PROTECTED
Write Cache Status        : Active, protected

AFA0>
```

RAID HW: afacli (3)

```
AFA0> disk list/full
Executing: disk list /full=TRUE

B:ID:L  Device Type     Removable media  Vendor-ID Product-ID        Rev   Blocks    Bytes/Block Usage     Shared Rate
------  --------------  ---------------  --------- ----------------  ----- --------- ----------- --------- ------ ----
0:00:0  Disk             N               HITACHI   DK32DJ-18MC        D4D4  35566478  512    Initialized    NO    160
0:01:0  Disk             N               HITACHI   DK32DJ-18MC        D4D4  35566478  512    Initialized    NO    160
1:02:0  Disk             N               HITACHI   DK32DJ-72MC        D4D4  143374650 512    Initialized    NO    160
1:03:0  Disk             N               HITACHI   DK32DJ-72MC        D4D4  143374650 512    Initialized    NO    160
1:04:0  Disk             N               HITACHI   DK32DJ-72MC        D4D4  143374650 512    Initialized    NO    160

AFA0> enclosure show temperature
Executing: enclosure show temperature

Enclosure
ID (B:ID:L) Sensor Temperature Threshold Status
----------- ------ ----------- --------- --------
 1  0:06:0   0       70 F   120  NORMAL
 1  0:06:0   1       77 F   120  NORMAL

AFA0> task list
Executing: task list

Controller Tasks

TaskId Function  Done%  Container State Specific1 Specific2
------ -------- ------- --------- ----- --------- ---------
No tasks currently running on controller

AFA0> close
Executing: close

FASTCMD> exit
Executing: exit

[root@afs4 tmp]#
```

RAID HW: megarc (1)

```
[root@afs4 dellsw]# megarc -AllAdpInfo

  ***************************************************************
       MEGARC PERC/CERC Configuration Utility(LINUX)-1.07(08-19-02)       By LSI Logic Corp.,USA
  ***************************************************************
  Type ? as command line arg for help

  AdapterNo  FirmwareType  CardType
  00       40LD/8SPAN    PERC 3/DC

[root@afs4 dellsw]# megarc -ctlrInfo -a0

  ***************************************************************
       MEGARC PERC/CERC Configuration Utility(LINUX)-1.07(08-19-02)       By LSI Logic Corp.,USA
  ***************************************************************
  Type ? as command line arg for help

  ***************************************************************
  Information of Adapter-0 (#Adapter(s) on system: 1)
  ***************************************************************

  Firmware Version : 161J  BIOS Version : 3.17
  Logical Drives : 09   DRAM : 128MB
  Rebuild Rate : 30%
  Flush Interval : 4 secs
  Number of Chnls : 2  Bios Status : Enabled
  Alarm state : Enabled   Auto rebuild : Enabled
  FW : SPAN-8,40-LD  BIOS Config AutoSelection : USER
  BIOS echos mesg : ON  BIOS stops on error : ON
  Initiator Id : 7(Clustered Firmware)
  ***************************************************************
[root@afs4 dellsw]#
```

RAID HW: megarc (2)

```
[root@afs4 dellsw]# megarc -dispCfg -a0

  ********************************************************************************************
       MEGARC PERC/CERC Configuration Utility(LINUX)-1.07(08-19-02) By LSI Logic Corp.,USA
  ********************************************************************************************
  Type ? as command line arg for help

  Finding Devices On Each PERC Adapter...
  Scanning Ha 0, Chnl 1 Target 15

  Error: Failed to get DiskCapacity: Ch 0, Id 0 on Adp-0

  Request Sense Data: 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00
                      00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00

  This disk is considered UNUSABLE.

  ********************************************************************************************
       Existing Logical Drive Informations By LSI Logic Corp.,USA
  ********************************************************************************************

    Logical Drive : 2( Adapter: 0 ):  Status: OPTIMAL
    ------------------------------------------------
          SpanDepth :01     RaidLevel: 0  RdAhead : Adaptive  Cache: DirectIo
          StripSz   :064KB   Stripes  : 1  WrPolicy: WriteBack

  Logical Drive 2 : SpanLevel_0 Disks
  Chnl  Target  StartBlock   Blocks      Physical Target Status
  ----  ------  ----------   ------      ----------------------
   0      03    0x00000000   0x0887b000   ?

  -1 to quit: LogDrive to view: 0..8:
```

Disco rotto...

RAID HW: megarc (2)

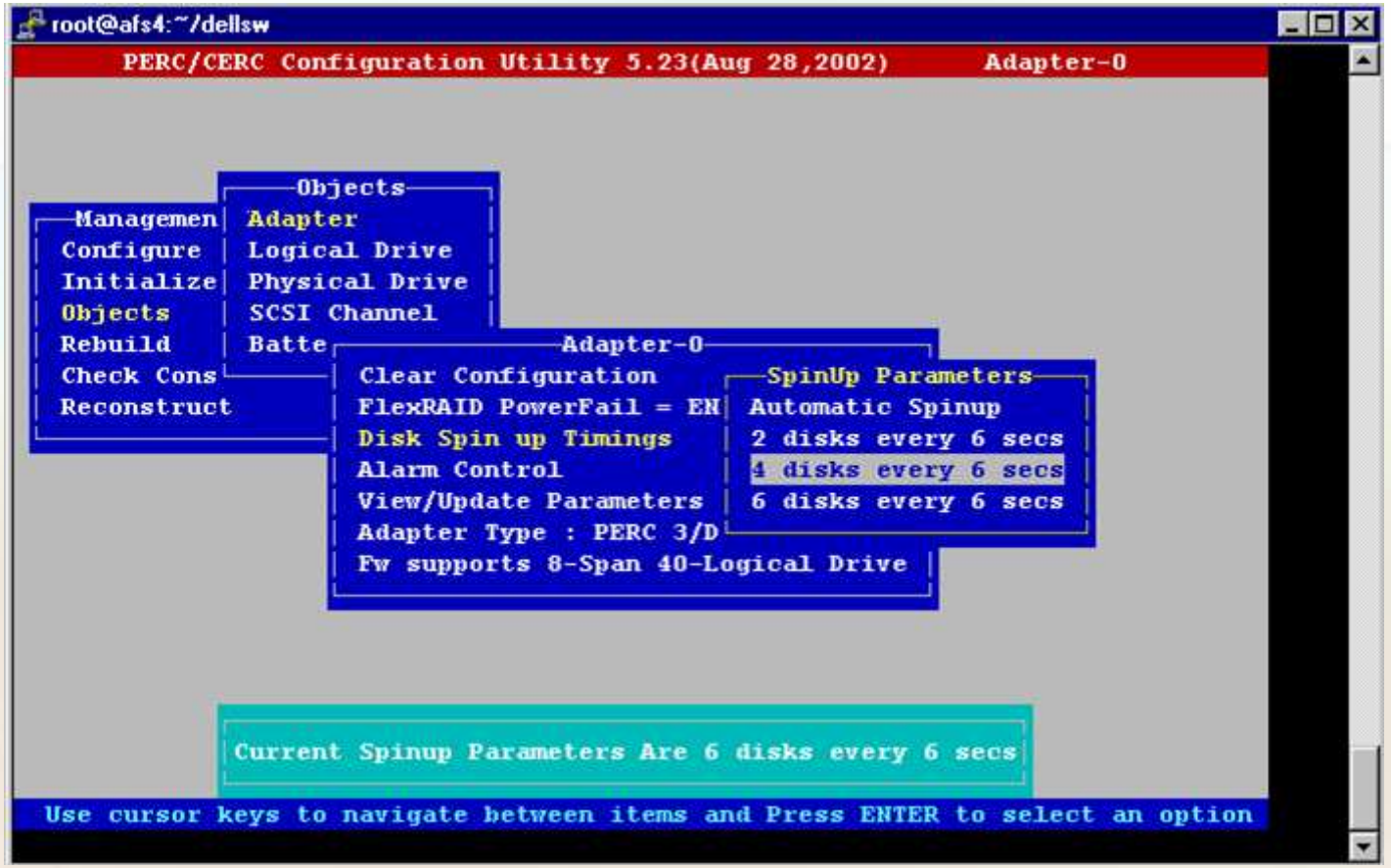
```
[root@afs4 dellsw]# megarc -ctlrInfo -a0

  ***************************************************************
       MEGARC PERC/CERC Configuration Utility(LINUX)-1.07(08-19-02)
       By LSI Logic Corp.,USA
  ***************************************************************
  Type ? as command line arg for help

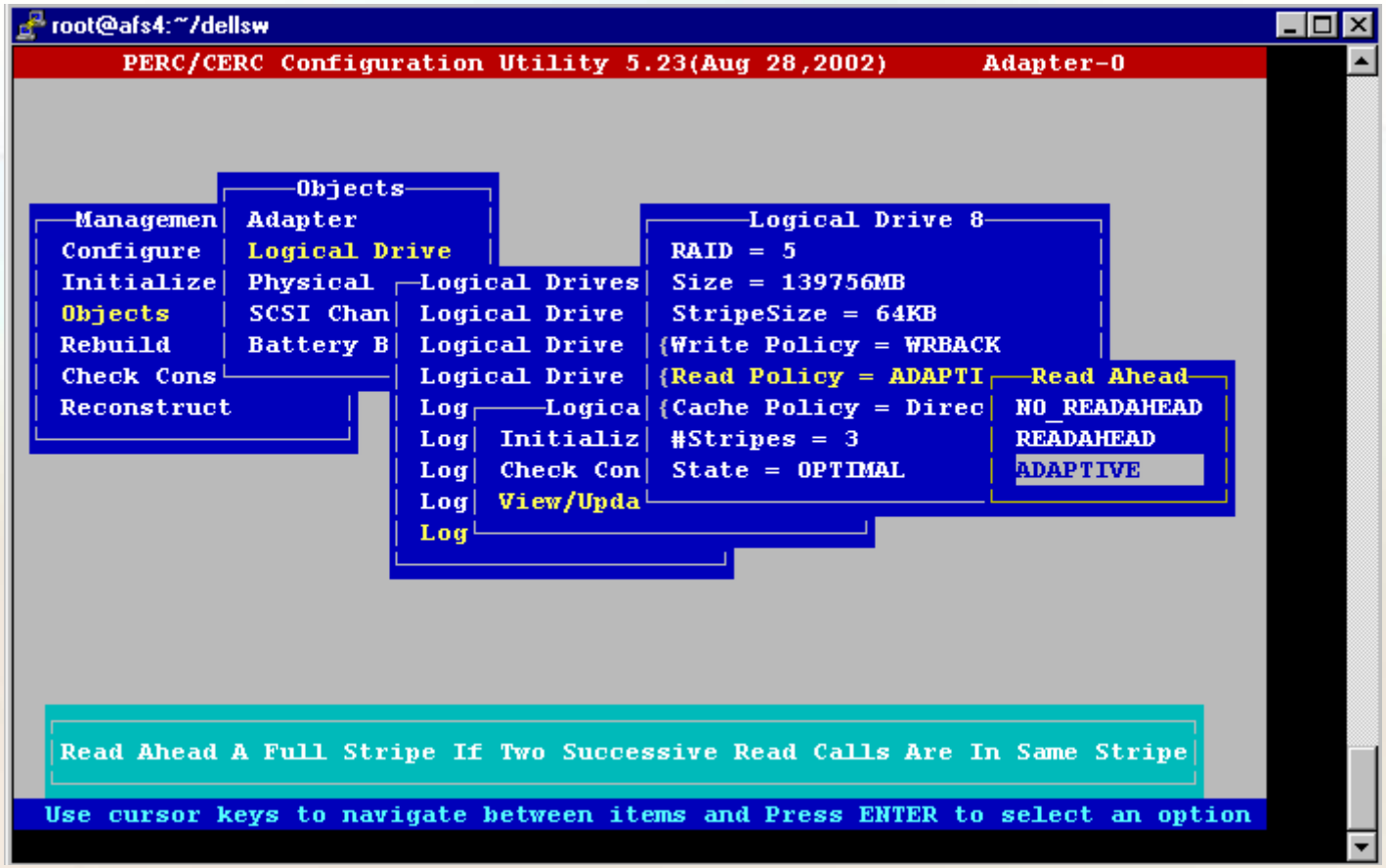
  ***************************************************************
  Information of Adapter-0 (#Adapter(s) on system: 1)
  ***************************************************************

  Firmware Version : 161J  BIOS Version : 3.17
  Logical Drives : 09   DRAM : 128MB
  Rebuild Rate : 30%
  Flush Interval : 4 secs
  Number of Chnls : 2  Bios Status : Enabled
  Alarm state : Enabled   Auto rebuild : Enabled
  FW : SPAN-8,40-LD  BIOS Config AutoSelection : USER
  BIOS echos mesg : ON  BIOS stops on error : ON
  Initiator Id : 7(Clustered Firmware)
  ***************************************************************
[root@afs4 dellsw]#
```

RAID HW: dellmgr



RAID HW: dellmgr



LVM: Logical Volume Manager

- Step necessari alla creazione di un Volume Group:
  - **Cambiare il tipo della partizione di disco da utilizzare nel volume group in 0x8E (Linux LVM partition)**
  - **Creare il volume fisico con “pvcreate <partizione> ...”**
  - **Creare il volume group con “vgcreae <nomevg> <partizione> …”**
- Creazione di un Logical Volume:
  - **create logical <size> <volumegroup>**

LVM: In pratica…(1)

- Lista delle partizioni/device di disco sul sistema:
  - ```
    [root@afs4 root]# lvmdiskscan
    lvmdiskscan -- reading all disks / partitions (this may take a while...)
    lvmdiskscan -- /dev/sda   [      68.24 GB] free whole disk
    lvmdiskscan -- /dev/sdb   [      68.24 GB] free whole disk
    lvmdiskscan -- /dev/sdc   [      68.24 GB] free whole disk
    lvmdiskscan -- /dev/sdd   [      68.24 GB] free whole disk
    lvmdiskscan -- /dev/sde   [      68.24 GB] free whole disk
    lvmdiskscan -- /dev/sdf   [      68.24 GB] free whole disk
    lvmdiskscan -- /dev/sdg   [      68.24 GB] free whole disk
    lvmdiskscan -- /dev/sdh   [     136.48 GB] free whole disk
    lvmdiskscan -- /dev/sdi   [      68.24 GB] free whole disk
    lvmdiskscan -- /dev/sdj   [     136.48 GB] free whole disk
    lvmdiskscan -- /dev/sdk1  [       9.77 GB] Primary  LINUX native partition [0x83]
    lvmdiskscan -- /dev/sdk2  [       1.46 GB] Primary  LINUX swap partition [0x82]
    lvmdiskscan -- /dev/sdk3  [       5.71 GB] Primary  LINUX native partition [0x83]
    lvmdiskscan -- /dev/sdl1  [     101.94 MB] Primary  LINUX native partition [0x83]
    lvmdiskscan -- /dev/md0   [     136.48 GB] free meta device
    lvmdiskscan -- /dev/md1   [      68.24 GB] free meta device
    lvmdiskscan -- /dev/md2   [     136.48 GB] free meta device
    lvmdiskscan -- 12 disks
    lvmdiskscan -- 10 whole disks
    lvmdiskscan -- 0 loop devices
    lvmdiskscan -- 3 multiple devices
    lvmdiskscan -- 0 network block devices
    lvmdiskscan -- 4 partitions
    lvmdiskscan -- 0 LVM physical volume partitions
    ```
    ```

LVM: In pratica…(2)

- Creare le partizioni di tipo “Linux LVM” sui device da utilizzare

- Creazione di physical volumes:
  - ```
    [root@afs4 root]# pvcreate  /dev/sdl1 /dev/sdl2
    pvcreate -- physical volume "/dev/sdl1" successfully created
    pvcreate -- physical volume "/dev/sdl2" successfully created
    ```

- Creazione del volume group “mastervg”:
  - ```
    [root@afs4 root]# vgcreate mastervg /dev/sdl1 /dev/sdl2
    vgcreate -- INFO: using default physical extent size 4 MB
    vgcreate -- INFO: maximum logical volume size is 255.99 Gigabyte
    vgcreate -- doing automatic backup of volume group "mastervg”
    vgcreate -- volume group "mastervg" successfully created and activated
    ```

- Creazione dei logical volume:
  - ```
    [root@afs4 root]# lvcreate -L 16384 mastervg
    lvcreate -- doing automatic backup of "mastervg”
    lvcreate -- logical volume "/dev/mastervg/lvol1" successfully created
    [root@afs4 root]# lvcreate -L 16384 mastervg
    lvcreate -- doing automatic backup of "mastervg”
    lvcreate -- logical volume "/dev/mastervg/lvol2" successfully created
    ```
    ```

LVM: In pratica…(3)

- E’ possibile verificare le caratteristiche di un volume group o dei volumi (logici e fisici) tramite:
  - **pvdisplay, vgdisplay, lvdisplay**
  - **Esempi:**
    - **[root@afs4 root]# pvdisplay /dev/sdk1**

```
--- Physical volume ---
PV Name               /dev/sdk1                     VG Name               mastervg
PV#                   1                 PV Status           available
Allocatable           yes (but full)                Cur LV                3
PE Size (KByte)       4096              Total PE            8747
Free PE               0                             Allocated PE          8747
PV Size               34.17 GB [71665902 secs] / NOT usable 4.19 MB [LVM: 162 KB]
PV UUID               PQfy2a-Yp3H-rEKk-LGfz-AKGP-HIeS-dGfMrZ
```

    - **[root@afs4 root]# vgdisplay mastervg**

```
--- Volume group ---
VG Name               mastervg                      VG Access             read/write
VG Status             available/resizable VG #              0
MAX LV                256               Cur LV              4
Open LV               0                             MAX LV Size           255.99 GB
Max PV                256               Cur PV              2
Act PV                2                             VG Size               68.34 GB
PE Size               4 MB              Total PE            17494
Alloc PE / Size       16384 / 64 GB     Free                PE / Size             1110 / 4.34 GB
VG UUID               Pv9LEN-7jVK-U2Cv-zZif-nkOx-Z5fY-pZlZyn
```

    - **[root@afs4 root]# lvdisplay /dev/mastervg/lvol1**

```
--- Logical volume ---
LV Name                /dev/mastervg/lvol1           VG Name                mastervg
LV Write Access        read/write                    LV Status              available
LV #                   1                 # open              0
LV Size                16 GB                         Current LE             4096
Allocated LE           4096                          Allocation             next free
Read ahead sectors     10000                         Block device           58:0
```

LVM: In pratica…(4)

- E’ possibile trarre informazioni su LVM, Volume Groups, Logical e Physical Volume anche accedendo a /proc/lvm/…
  -
```
[root@afs4 root]# cat /proc/lvm/global
LVM module LVM version 1.0.3(19/02/2002)

Total:  1 VG  2 PVs  4 LVs (0 LVs open)
Global: 267884 bytes malloced   IOP version: 10   3 days 23:58:27 active

VG:  mastervg  [2 PV, 4 LV/0 open]  PE Size: 4096 KB
  Usage [KB/PE]: 71655424 /17494 total  67108864 /16384 used  4546560 /1110 free
  PVs: [AA] sdk1                 35827712 /8747    35827712 /8747           0 /0
       [AA] sdk2                 35827712 /8747    31281152 /7637     4546560 /1110
    LVs: [AWDL  ] lvol1                      16777216 /4096     close
         [AWDL  ] lvol2                      16777216 /4096     close
         [AWDL  ] lvol3                      16777216 /4096     close
         [AWDL  ] lvol4                      16777216 /4096     close
```
  -
```
[root@afs4 root]# cat /proc/lvm/VGs/mastervg/group
name:         mastervg
size:         71655424
access:       3
status:       5
number:       0
LV max:       256
LV current:   4
LV open:      0
PV max:       256
PV current:   2
PV active:    2
PE size:      4096
PE total:     17494
PE allocated: 16384
uuid:         Pv9L-EN7j-VKU2-CvzZ-ifnk-OxZ5-fYpZ-lZyn
```

LVM: In pratica…(5)

- I vari comandi che permettono di gestire LVM, i Volume Groups, i Logical e i Physical Volume utilizzano i dati contenuti in
  - /etc/lvmtab (contiene la lista dei Volume Groups)
  - /etc/lvmtab.d/<VGname>
- Tali file sono ricostruiti al boot tramite il comando “vgscan”
- E’ possibile utilizzare vgscan anche per forzare la ricostruzione dei database relativi ad LVM

LVM: In pratica…(6)

- Altri comandi per la gestione di LVM sono:
  - `lvchange` change attributes of a logical volume
  - `lvcreate` create a logical volume in an existing volume group
  - `lvdisplay` display attributes of a logical volume
  - `lvextend` extend the size of a logical volume
  - `lvreduce` reduce the size of a logical volume
  - `lvremove` remove a logical volume
  - `lvrename` rename a logical volume
  - `lvscan` scan (all disks) for logical volumes
  - `pvchange` change attributes of a physical volume
  - `pvdata` shows debugging information about a physical volume
  - `pvdisplay` display attributes of a physical volume
  - `pvscan` scan all disks for physical volumes
  - `vgcfgbackup` backup volume group descriptor area
  - `vgcfgrestore` restore volume group descriptor area
  - `vgchange` change attributes of a volume group
  - `vgck` check volume group descriptor area
  - `vgcreate` create a volume group
  - `vgdisplay` display attributes of volume groups
  - `vgexport` make volume groups unknown to the system
  - `vgextend` add physical volumes to a volume group
  - `vgimport` make volume groups known to the system
  - `vgmerge` merge two volume groups
  - `vgmknodes` create volume group directory and special files
  - `vgreduce` reduce a volume group
  - `vgremove` remove a volume group
  - `vgrename` rename a volume group
  - `vgscan` scan disks for VGs and build /etc/lvmtab and /etc/lvmtab.d/*
  - `vgsplit` split a volume groups

File System – Generalita’

- File: Spazio logico contiguo di dati
- File System: Organizza lo spazio fisico per poter contenere e permettere di operare sui file (tramite FCB)
  - FCB: File Control Block
    - Insieme di informazioni necessarie a descrivere e ad accedere un file
  - L’organizzazione o allocazione dello spazio fisico puo’ essere:
    - Contigua
    - “Linked” (Lista concatenata)
    - “Indexed”

File System - Organizzazione

- Contigua
  - Un file e' individuato tramite posizione fisica del suo inizio e la sua dimensione
  - Random access, spreco di spazio, dimensione dei file non modificabile
- "Linked" (Lista concatenata)
  - Organizzazione a blocchi (di dati)
  - Un file e' individuato dall'indirizzo del suo blocco iniziale, il quale contiene come in una lista concatenata anche l'eventuale indirizzo del blocco successivo
  - Solo accesso sequenziale, minimo spreco di spazio
  - Es.: FAT (File Allocation Table)
- "Indexed"
  - Organizzazione a blocchi (di dati e indici)
  - Un file e' individuato tramite un "index block" (i-node) che contiene i puntatori a tutti i blocchi fisici con i dati
  - L'i-node contiene altre informazioni quali il tipo di file, le modalita' di accesso e ownership, le dimensioni, le informazioni su date di creazione modifica e accesso (quindi FCB = i-node)
  - Random access, minimo spreco di spazio
  - Nella sua naturale evoluzione i file system Indexed strutturano gli i-node in piu' parti:
    - **indicizzamento diretto ai blocchi dati**
    - **indicizzamento indiretto ai blocchi dati (tramite un ulteriore "index block")**
    - **Indicizzamento indiretto doppio e triplo ai blocchi dati (tramite due o tre livelli di"index block")**

File System - Linux

- I file system generalmente in uso sui sistemi Unix e Linux in particolare sono di tipo ad allocazione indexed
- Uno speciale blocco detto “Superblock” contiene i dati vitali del file system
  - Il numero di blocchi e i-node totali
  - Il numero di blocchi e i-node liberi
  - Una lista di indici di i-node liberi
  - Puntatore al primo blocco libero
  - Etc.
- Il Superblock e’ mantenuto in memoria dal kernel che lo aggiorna sul disco ogni 30 secondi circa
- Per sicurezza il Superblock e mantenuto in piu’ copie sul disco

File System – mount e umount

- Un file system e' disponibile al sistema e alle applicazioni solo se viene "montato" (comando mount)
  - Il comando di mount permette anche di modificare le modalita' di accesso al file system (ad es.: solo accesso read-only, no execute permission, etc.)
- In caso di "kernel panic", guasto hardware o mancanza improvvisa di corrente elettrica le operazioni di scrittura sui dischi possono essere interrotte prima che queste vengano portate a termine e percio' il file system potrebbe essere inconsistente nelle sue varie strutture (Superblock, i-node, blocchi dati).
  Per questo motivo ad ogni restart il sistema controlla un flag sul file system al fine di verificare la corretta terminazione delle operazioni di scrittura e di successivo umount.
- Se questo flag indica una mancata operazione di umount (not clean) e' necessario eseguire un file system check (comando fsck), questo puo' avvenire automaticamente o manualmente (dipende dalla configurazione dei flag di mount)

File System – fsck

- L’operazione di file system check e’ suddivisa in passaggi:
  - Pass 1: Checking inodes, blocks, and sizes
  - Pass 2: Checking directory structure
  - Pass 3: Checking directory connectivity
  - Pass 4: Checking reference counts
  - Pass 5: Checking group summary information

File System – Journal

- Su file system di grandi dimensioni o con un grande numero di file l'operazione di file system check puo' richiedere tempi molto lunghi
- Per ovviare a tale inconveniente possono essere utilizzati dei file system di tipo journaled
- Il "Journal" e' un file speciale che contiene la traccia (log) delle operazioni di scrittura che vengono eseguite sul file system (modifica di i-node, mappe di allocazione, etc.)
- In caso di "dirty" file system viene eseguito un "replay" del journal al fine di verificare le operazioni di scrittura e riportare il file system in stato consistente
- Il Journal puo' diventare il collo di bottiglia delle operazioni di scrittura
  - Per evitare cio' puo' essere utile definire il journal su un device separato

File System – Ext2

- Ottimizzato per diminuire la frammentazione:
- Usa piccoli blocchi (4K)
- Lo spazio disco e' ripartito in gruppi di blocchi contigui, allocando blocchi logici consecutivi di un file in blocchi fisici contigui
- "Localizzazione": cerca di tenere le directory con i relativi i-node e i blocchi dei corrispondenti file nello stesso gruppo (diminuisce il seek time)
- Possibilmente le directory vengono messe in gruppi diversi
- Alcuni dati:
  - Max filename size: 255 bytes
  - Max filesize: 2 TBytes
  - Max file system size: 32 TBytes

File System – Ext3

- Estensione dell'Ext2 di cui mantiene la stessa struttura dei metadata
- Journal con 3 possibili modalita':
  - Journal: mantiene sia i metadata che i dati. Piu' affidabile ma piu' lento (deve scrivere 2 volte)
  - Writeback: mantiene solo i metadata. Piu' veloce, ma in caso di crash durante l'append di un file puo' portare il file a contenere dati non validi nella parte finale del file
  - Ordered: mantiene solo i metadata come nella modalita' Writeback, ma i dati sono scritti solo dopo l'aggiornamento dei metadata. Compromesso tra affidabilita' e velocita'. E' l'opzione di default
- Se il replay del Journal fallisce, viene eseguito un fsck standard
- Alcuni dati:
  - Max filename size: 255 bytes
  - Max filesize: 2 TBytes
  - Max file system size: 32 TBytes

File System – ReiserFS

- Metadata Journal
- Tabelle di i-node, directory e blocchi dati strutturate con Alberi Binari Bilanciati, Balanced Tree (B+)
- Possibilita’ di ingrandire il file system online (e’ possibile invece ridurlo solo offline)
- “Tail packing”
  - Schema per la riduzione della frammentazione interna dei file. Puo’ essere disattivato per migliorare le performance.
- Alcuni dati:
  - Max filename size: 255 bytes
  - Max filesize: 8 TBytes
  - Max file system size: 16 TBytes

File System – XFS

- Sviluppato da SGI per Irix e' stato rilasciato sotto la "GNU General Public License" nel 2000
- Metadata Journal con write-ahead
  - Basato su transazioni. File system rimane consistente dopo ogni transazione.
- Directory strutturate con un Albero Binario Bilanciato, Balanced Tree (B+)
  - I nomi dei file sono convertiti in un hash a 4 byte rendendo la ricerca velocissima
- Allocazione dinamica dei blocchi contenenti i-node
- Possibilita' di ingrandire il file system online (non e' possibile invece ridurlo)
- Deframmentazione online
- Real-time I/O API (adatto a applicazioni real-time applications, come il video streaming)
- "Allocate-on-flush" (riduce la frammentazione per i file che crescono lentamente)
- Creazione del file system molto veloce
- Alcuni dati:
  - Max filename size: 255 bytes
  - Max filesize: 9 EBytes
  - Max file system size: 9 EBytes

Creazione di File System

- Creato il device di disco (sia esso una partizione di un disco fisico, un RAID sw o hw, un logical volume) e' possibile creare su di esso un file system.
- Su Linux puo' essere usata la "suite" mkfs che richiamera' l'eseguibile associato al tipo di file system richiesto, i due comandi che seguono sono quindi equivalenti:
  - mkfs -t xfs /dev/sda1
  - mkfs.xfs /dev/sda1

Esempi di mkfs su Linux (ext2)

```
[root@afs4 tmp]# mkfs -t ext2 /dev/md0
mke2fs 1.27 (8-Mar-2002)
Filesystem label=
OS type: Linux
Block size=4096 (log=2)
Fragment size=4096 (log=2)
8945664 inodes, 17888752 blocks
894437 blocks (5.00%) reserved for the super user
First data block=0
546 block groups
32768 blocks per group, 32768 fragments per group
16384 inodes per group
Superblock backups stored on blocks:
  32768, 98304, 163840, 229376, 294912, 819200, 884736, 1605632, 2654208,
  4096000, 7962624, 11239424

Writing inode tables: done
Writing superblocks and filesystem accounting information: done

This filesystem will be automatically checked every 35 mounts or
180 days, whichever comes first.  Use tune2fs -c or -i to override.
```

Esempi di mkfs su Linux (ext3)

```
[root@afs4 tmp]# mkfs -t ext3 /dev/sdd1
mke2fs 1.27 (8-Mar-2002)
Filesystem label=
OS type: Linux
Block size=4096 (log=2)
Fragment size=4096 (log=2)
8945664 inodes, 17888369 blocks
894418 blocks (5.00%) reserved for the super user
First data block=0
546 block groups
32768 blocks per group, 32768 fragments per group
16384 inodes per group
Superblock backups stored on blocks:
  32768, 98304, 163840, 229376, 294912, 819200, 884736, 1605632,
2654208, 4096000, 7962624, 11239424

Writing inode tables: done
Creating journal (8192 blocks): done
Writing superblocks and filesystem accounting information: done

This filesystem will be automatically checked every 28 mounts or
180 days, whichever comes first.  Use tune2fs -c or -i to override.
```

Esempi di mkfs su Linux (reiserfs)

- ```
  [root@afs4 tmp]# mkfs -t reiserfs /dev/md0
  <-------------mkfs.reiserfs, 2002------------>
  reiserfsprogs 3.6.2
  mkfs.reiserfs: Guessing about desired format..
  mkfs.reiserfs: Kernel 2.4.19-SGI_XFS_1.2pre4smp is running.
  Format 3.6 with standard journal
  Count of blocks on the device: 17888752
  Number of blocks consumed by mkreiserfs formatting process: 8757
  Blocksize: 4096
  Hash function used to sort names: "r5”
  Journal Size 8193 blocks (first block 18)
  Journal Max transaction length 1024
  inode generation number: 0
  UUID: b399ac6b-d6e5-47e6-9ce5-f02969ca2cb6
  ATTENTION: YOU SHOULD REBOOT AFTER FDISK! ALL DATA WILL BE LOST ON '/dev/md0'!
  Continue (y/n):y
  Initializing journal - 0%....20%....40%....60%....80%....100% Syncing..ok

  The Defense Advanced Research Projects Agency (DARPA) is the primary sponsor of Reiser4. DARPA does not endorse this project; it merely sponsors it.
  Continuing core development of version 3 is mostly paid for by Hans Reiser from money made selling licenses in addition to the GPL to companies who don't
  want it known that they use ReiserFS as a foundation for their proprietary product.
  And my lawyer asked 'People pay you money for this?'.  Yup.  Hee Hee.  Life is good.  If you buy ReiserFS, you can focus on your value add rather than
  reinventing an entire FS.  You should buy some free software too.…
  SuSE pays for continuing work on journaling for version 3, and paid for muchof the previous version 3 work.  Reiserfs integration in their distro is
  consistently solid.
  MP3.com paid for initial journaling development.
  Bigstorage.com contributes to our general fund every month, and has done so for quite a long time.
  Thanks to all of those sponsors, including the secret ones.  Without you, Hans would still have that day job, and the merry band of hackers would be
  missing quite a few.…
  Have fun.
  ```
  ```

Esempi di mkfs su Linux (xfs)

- ```
  [root@afs4 tmp]# mkfs -t xfs /dev/md2
  meta-data=/dev/md2               isize=256    agcount=35, agsize=1048576 blks
  data     =                       bsize=4096   blocks=35777504, imaxpct=25
           =                       sunit=16     swidth=48 blks, unwritten=0
  naming   =version 2              bsize=4096
  log      =internal log           bsize=4096   blocks=4368, version=1
           =                       sunit=16 blks
  realtime =none                   extsz=196608 blocks=0, rtextents=0
  ```

- E’ possibile rivedere tali parametri con il comando xfs_info:

  - ```
    [root@afs4 tmp]# mount /dev/md2 /mnt/xfs
    [root@afs4 tmp]# xfs_info /mnt/xfs
    meta-data=/mnt/xfs               isize=256    agcount=35, agsize=1048576 blks
    data     =                       bsize=4096   blocks=35777504, imaxpct=25
             =                       sunit=16     swidth=48 blks, unwritten=0
    naming   =version 2              bsize=4096
    log      =internal               bsize=4096   blocks=4368 version=1
             =                       sunit=0 blks
    realtime =none                   extsz=196608 blocks=0, rtextents=0
    ```

Esempi di fsck su Linux (ext2)

```
# fsck.ext2 -f ./testfs
e2fsck 1.34 (25-Jul-2003)
Pass 1: Checking inodes, blocks, and sizes
Inode 11 has illegal block(s).  Clear<y>? yes

Illegal block #4 (167772426) in inode 11.  CLEARED.
Inode 11, i_blocks is 24, should be 22.  Fix<y>? yes

Pass 2: Checking directory structure
Directory inode 11 has an unallocated block #4.  Allocate<y>? yes

Pass 3: Checking directory connectivity
Pass 4: Checking reference counts
Pass 5: Checking group summary information

./testfs: ***** FILE SYSTEM WAS MODIFIED *****
./testfs: 11/8192 files (9.1% non-contiguous), 1052/32764 blocks
```

Esempi di fsck su Linux (ext3)

```
# fsck.ext3 -f ./testfs
e2fsck 1.34 (25-Jul-2003)
Pass 1: Checking inodes, blocks, and sizes
node 8 has illegal block(s).  Clear<y>? Yes
Illegal block #1011 (184549386) in inode 8.  CLEARED.
Illegal block #1012 (201326597) in inode 8.  CLEARED.
… omissis …
Illegal block #1021 (352321541) in inode 8.  CLEARED.
Too many illegal blocks in inode 8.
Clear inode<y>? Yes
Restarting e2fsck from the beginning...
Superblock has a bad ext3 journal (inode 8).
Clear<y>? Yes
*** ext3 journal has been deleted - filesystem is now ext2 only ***
Pass 1: Checking inodes, blocks, and sizes
Journal inode is not in use, but contains data.  Clear<y>? Yes
Pass 2: Checking directory structure
Pass 3: Checking directory connectivity
Pass 4: Checking reference counts
Pass 5: Checking group summary information
Block bitmap differences:  -(274--4387) +(8445--8452) +(16635--16642) +(24829--24836) -32760
Fix<y>?
Free blocks count wrong for group #0 (3805, counted=7919).
./testfs: ***** FILE SYSTEM WAS MODIFIED *****
./testfs: 11/8192 files (0.0% non-contiguous), 1052/32768 blocks
```

Benchmarking: breve introduzione

- Per valutare la bonta' di un sistema destinato alla distribuzione di risorse disco esistono diversi software che eseguono vari tipi di accessi al disco
- E' comunque possibile avere un'idea delle prestazioni anche utilizzando semplicemente il comando "dd"
  - dd if=/dev/zero of=/mnt/testfs/ddtest bs=1024k count=1024
  - dd if=/mnt/testfs/ddtest of=/dev/null bs=64k count=2048
- Nel primo dd scriviamo 1Gb a blocchi di 1Mb sul file system scelto, nel secondo leggiamo 256Mb a blocchi di 64Kb
- L'uso di /dev/zero e /dev/null permette produrre o ricevere dati senza interagire con altri dispositivi fisici limitando quindi l'influenza di operazioni diverse da quelle che vogliamo valutare

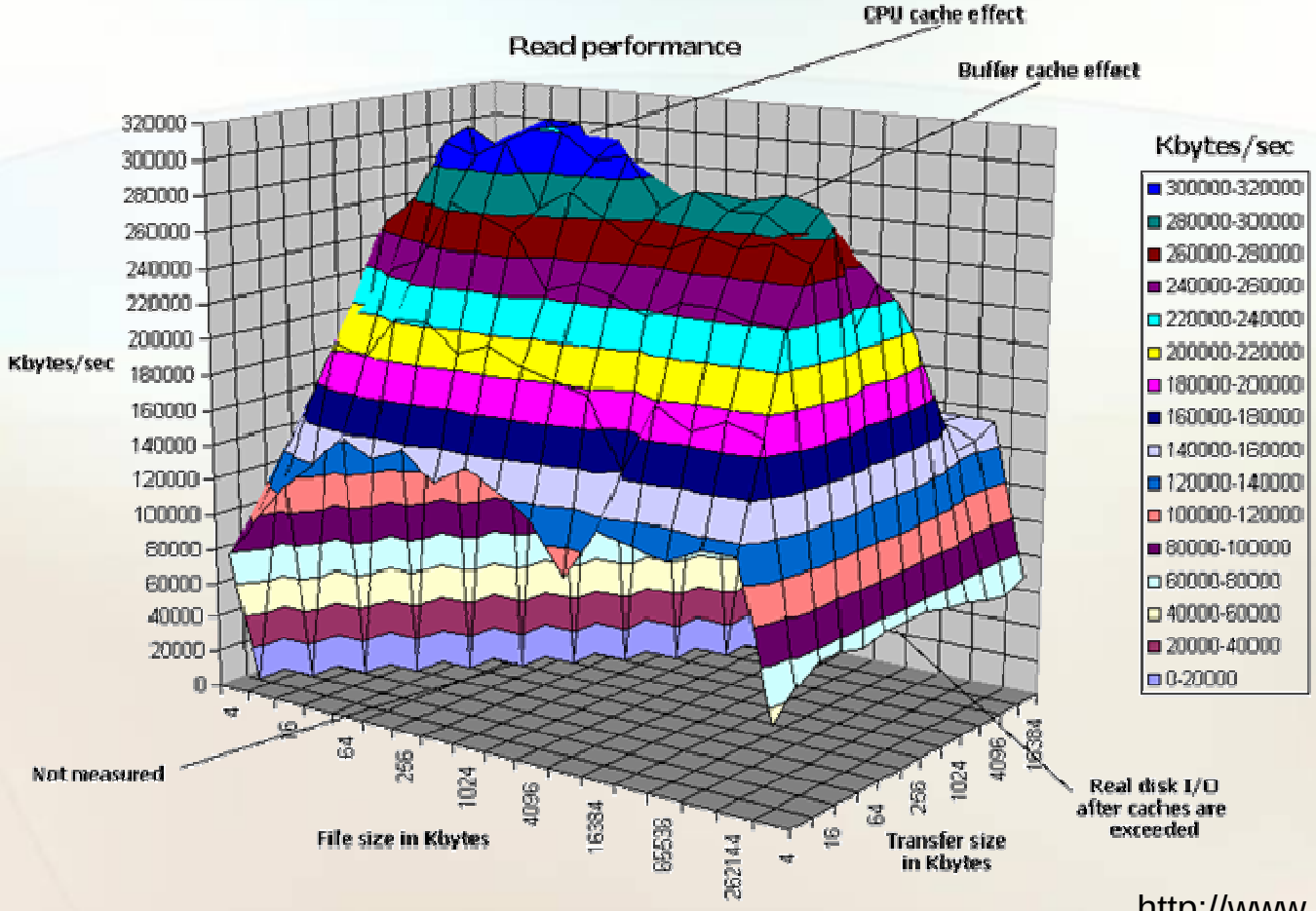
Benchmarking: breve introduzione

- Il semplice “dd” pero’ non mette in evidenza l’influenza determinata sulle operazioni di I/O da parte della Memory Cache utilizzata dal sistema; un modo semplice per rendere percebile cio’ e’ l’uso del comando “sync” subito dopo la conclusione del “dd”
  - dd if=/dev/zero of=/mnt/testfs/ddtest bs=1024k count=1024; sync
- Se il sistema e’ previsto per un uso facilmente “simulabile” con “dd” un suo uso accurato unito al “sync” puo’ dare una prima stima delle prestazioni del sistema
- Per risultati piu’ accurati si possono usare sw specifici (ne esistoni sia di gratuiti che a pagamento), come ad esempio:
  - IOzone
  - Bonnie, Bonnie++
  - AIM7
  - SpecSFS

Benchmarking - IOzone

- http://www.iozone.org
- Numerosi tipi di test:
  - Write, Re-write, Record Rewrite
  - Read, Re-Read, Backwards Read
  - Random Read, Random Write, Random Mix (Alternanza di operazioni di lettura e scrittura)
  - Strided Read (Alternanza di operazioni di lettura e seek in avanti)
  - Fwrite, Frewrite, Fread, Freread (utilizzo di fwrite() e fread() che eseguono operazioni bufferizzate)
- Possibilita' di flusso singolo o flussi multipli
- Test con processi multipli
- Possibilita' di produrre file Excel per la generazione di grafici

Benchmarking - IOzone



http://www.iozone.org

Fine Prima Parte

Sandro.Angius@lnf.infn.it