Dall'hard disk al file system distribuito
(un mix tecnologico)

Andrei Maslennikov
CASPUR / INFN

8 maggio 2007 - Rimini

Sommario

- Dischi, interfacce, RAID
- Indagine CERN sulle “corruzioni silenziose”
- Notizie FS (locali e distribuiti)
- Situazione FSD nell’HEP e oltre (HEPiX FSWG)

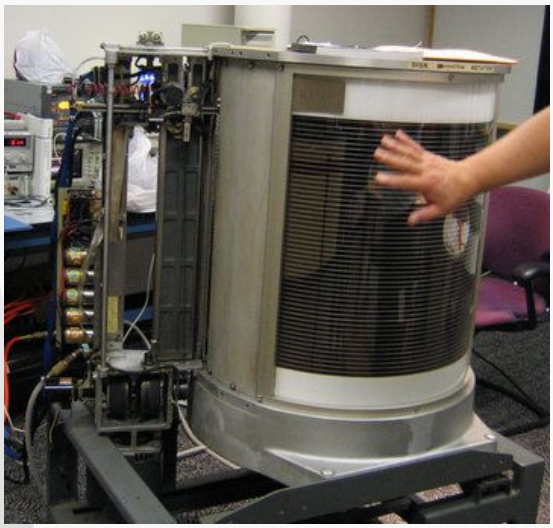
2

Dischi, interfacce, RAID...

Dischi: sempre più capienti...

Primo hard disk: IBM RAMAC 350 (1956)

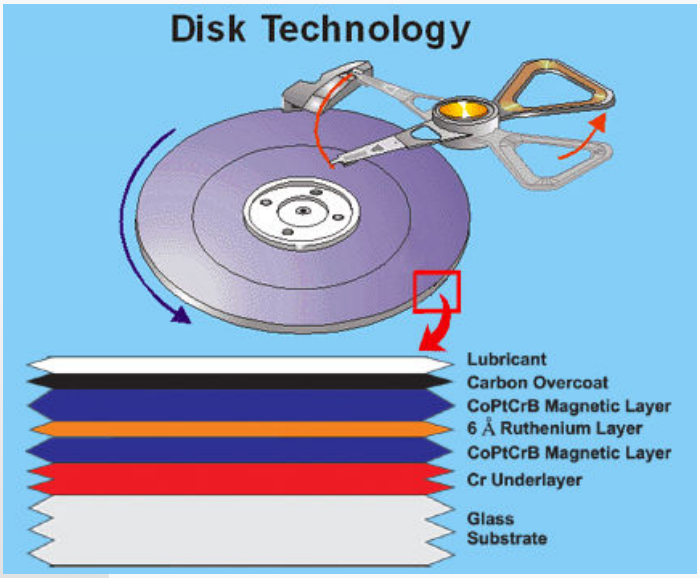
- 1 testina
- 50 platter magnetici del diametro di un metro
- tempo medio d'accesso – 75 millisecondi
- capacità: 5 MB



Il principio è rimasto invariato da quell'epoca...

Dischi, interfacce, RAID

I primi platter venivano coperti con la vernice usata per rivestire il Golden Gate Bridge. Ecco invece come si presentano le superfici di un platter moderno:



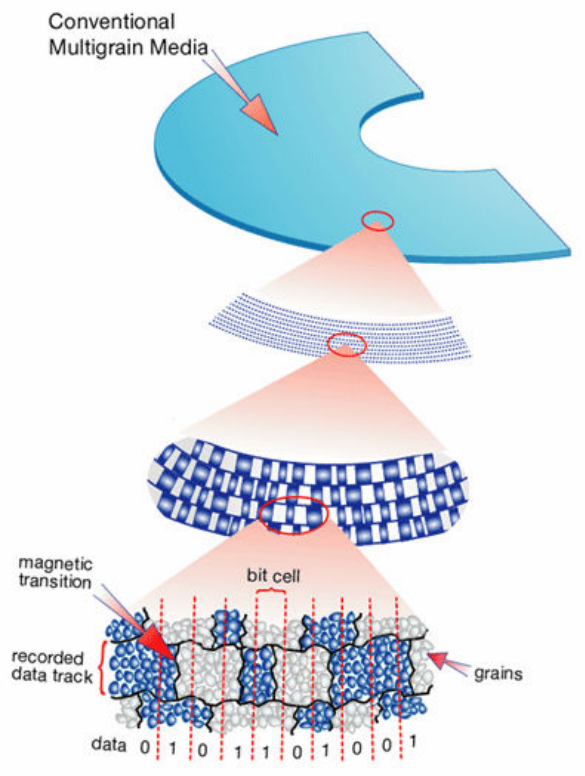
La capacità di un platter è cresciuta moltissimo nel tempo e sta aumentando ancora. Ciò è possibile solo grazie alle nuove tecnologie di magnetizzazione (quella tradizionale ha quasi raggiunto il limite imposto dal Super Paramagnetic Limit).

5

Dischi, interfacce, RAID

Tecnica tradizionale ("Longitudinal"):

- Ogni traccia racchiude le "bit cells" contenenti 50-100 granuli magnetici ciascuna
- L' "1" è individuato dalla presenza di una boundary tra le regioni di opposta magnetizzazione all'interno di una "bit cell"
- Lo "0" è invece individuato da un'area senza boundary



6

Dischi, interfacce, RAID

Aggirare l'SPL!

- I platter tradizionali disponibili oggi in commercio hanno la densità < 80 gigabit/inch^2

- Super Paramagnetic Limit:
  se V è il volume di un granulo e Ku la sua energia di anisotropia magnetica
  e il prodotto V*Ku risulta essere inferiore a un certo valore denominato
  Super Paramagnetic Limit (SPL) allora la magnetizzazione di questo granulo può
  cambiare spontaneamente stato (flip).

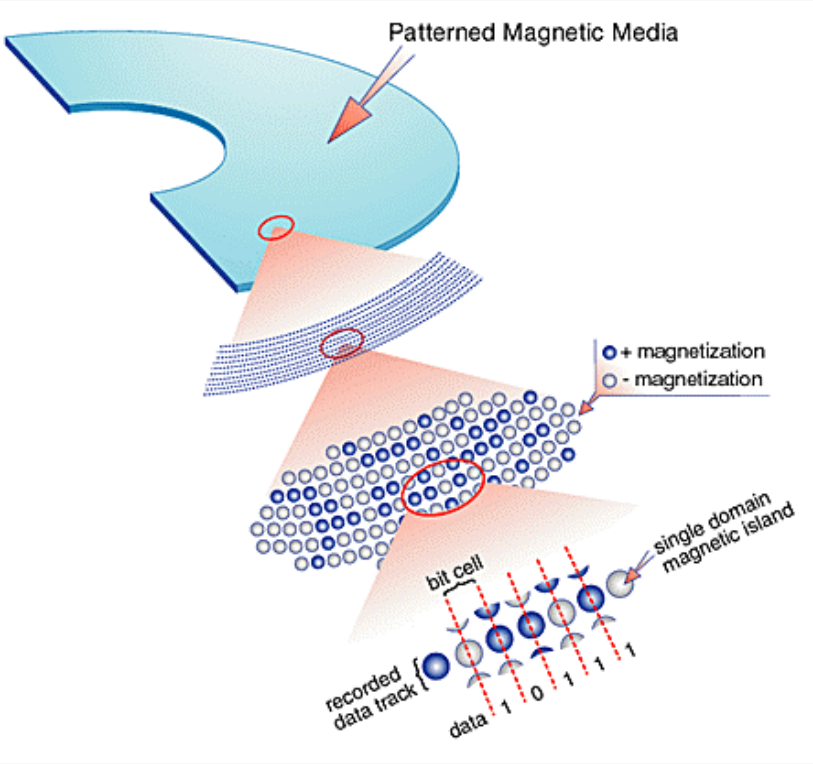
- Per aumentare la capacità di un platter bisogna diminuire le dimensioni
  dei granuli. Ma nello stesso tempo è necessario aumentare Ku per
  non raggiungere lo SPL stimato oggi al di sopra dei 100-200 gigabit/inch^2.
  La capacità di magnetizzazione delle testine ha un limite…

- Soluzioni:
  - Patterned Media
  - Perpendicular Recording
  - Heat Assisted Magnetic Recording

7

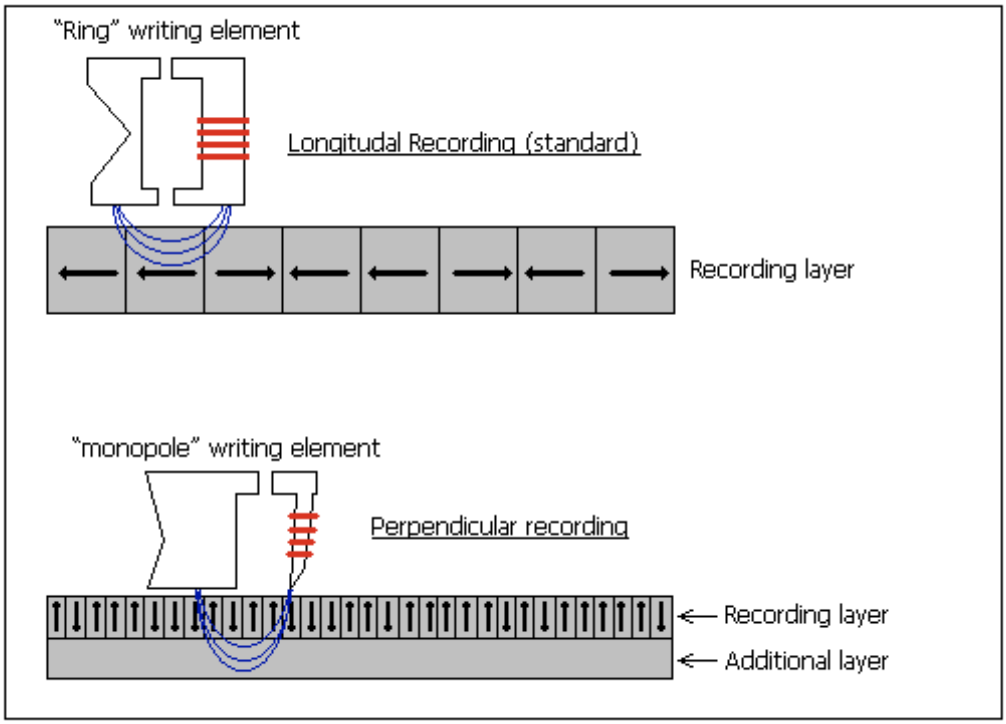
Dischi, interfacce, RAID

Patterned Media (Hitachi): la magnetizzazione avviene a livello di granulo



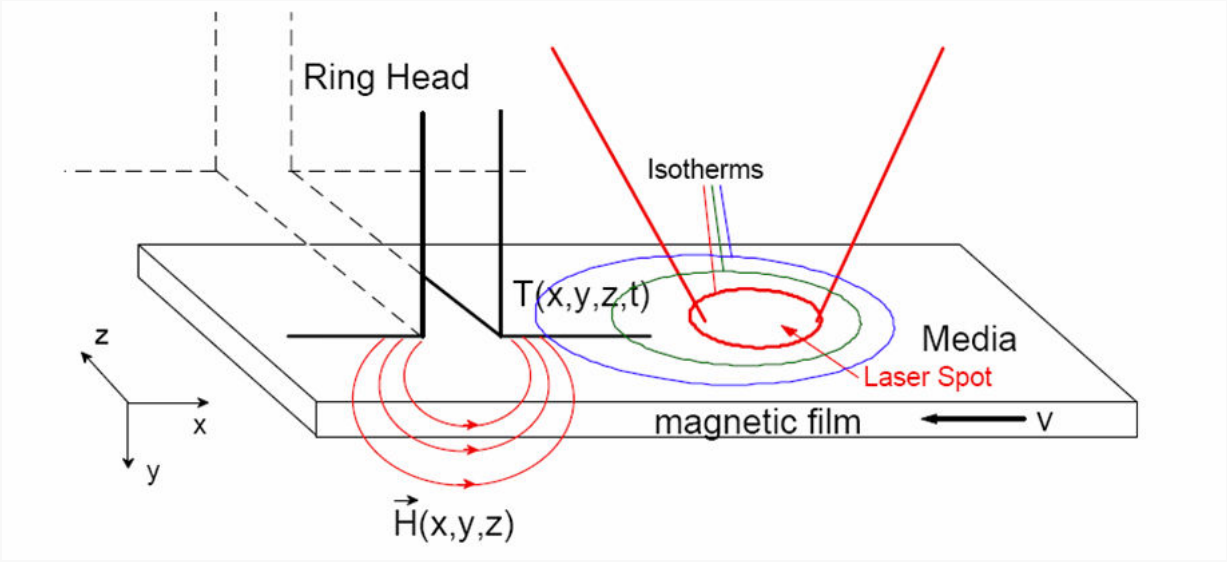
Dischi, interfacce, RAID

Perpendicular Recording



Dischi, interfacce, RAID

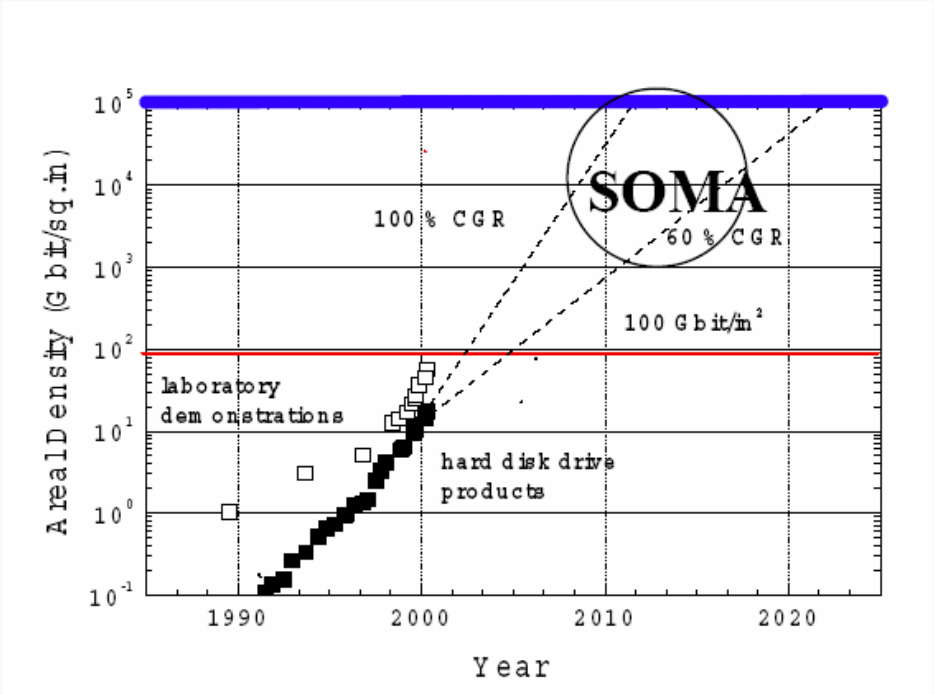
HAMR – Heat Assisted Magnetic Recording (=OAMR, =TAMR)



Dischi, interfacce, RAID

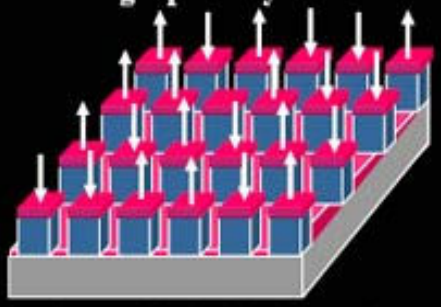
HAMR+SOMA (Self Organized Magnetic Arrays) ► 2010

Con i granuli FePt da 6 nm si potrà potenzialmente raggiungere la densità di **50 terabit/inch²**.

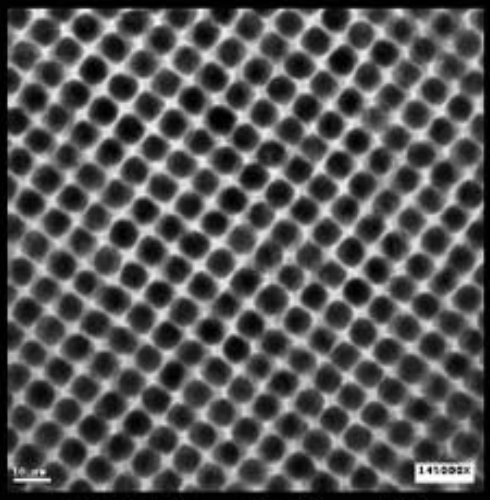


Bit Patterned Media
Lithography vs Self Organization

Lithographically Defined



FePt SOMA media



- Major obstacle is finding low cost means of making media.
  - At 1 Tbpsi, assuming a square bit cell and equal lines and spaces, 12.5 nm lithography would be required.
  - Semiconductor Industry Association roadmap does not provide such linewidths within the next decade.

- 6.3+/-0.3 nm FePt particles
- $\sigma_{Diameter} \cong 0.05$

S. Sun, Ch. Murray, D. Weller, L. Folks, A. Moser, Science 287, 1989 (2000).

Dischi, interfacce, RAID

I trend della tecnologia del disco

- **Dischi fissi da 3.5 pollici**, Gennaio 2003 – Gennaio 2007. **La velocità di rotazione rimane invariata**, la capacità è cresciuta di un fattore 2-4 (Hitachi, Seagate):

Velocità di rotazione	Gennaio 2003	Gennaio 2007	
15 KRPM	73GB, 3.3ms	300 GB, 3.5ms	(2.70 Euro/GB)
10 KRPM	182GB, 4.7ms	300 GB, 4.3ms	
7.2 KRPM	300GB, 9.3ms	1000 GB, 9.2ms	(0.45 Euro/GB)

13

Dischi, interfacce, RAID

- **SSD (Solid State Disks):** sono ancora molto costosi e vengono impegnati soprattutto in occasioni speciali.

  Hanno una bassissima latenza, un'altissima velocità di accesso ai dati e ovviamente godono della totale assenza di complessità della meccanica ad alta precisione.

  Oggi sono disponibili i device con capacità fino a 128 GB (Samsung, Sandisk, A-DATA, SimpleTech). Costi: sono ancora proibitivi, il costo di un singolo GB è spesso cento volte superiore di un GB di un disco tradizionale.

  Un'esempio: sistema Solid State RamSan-400 di Texas Memory Systems:

- Fino a 400 KIOPS, 3 GB/sec random sustained external throughput
- Ottimo per velocizzare i database

Dischi, interfacce, RAID

Ci servono veramente i dischi di grosso taglio?

- La risposta è "**NI**":

  SI: per il Fast Nearline Tier, per le applicazioni deep archiving ecc.

  NO: per l'intensive random I/O abbiamo bisogno di mantenere bilanciati la somma di IOPS proveniente da tutti gli stream e il numero di spindle, in modo da raggiungere un numero massimale di MB/spindle

- Specificare un limite in MB per spindle significherebbe sottoutilizzare i costosi slot dei sistemi RAID.

- L'ottimizzazione intelligente dei costi è comunque possibile. Una strada da considerare potrebbe essere quella della spartizione logica dei logical drive basati su pochi dischi di grande capacità. Destinando una parte di un grande logical drive a I/O esigente ed un'altra al nearline tier, si riesce a sfruttare al meglio i dischi di grosso taglio.

15

Dischi, interfacce, RAID

Due parole sulle interfacce

- Parallele o seriali? Il bus parallelo ha quasi definitivamente perso terreno, lasciando sempre più spazio a quello seriale. Ciò è dovuto da una parte ai limiti dell'architettura parallela (signal skew, cross-talks, ecc.), dall'altra al fatto che i bus seriali di oggi possono operare ad altissime frequenze.

- Interfacce disponibili al giorno d'oggi:

Parallele :

- ATA / IDE/ EIDE (Advanced Technology Attachment / Integrated Drive Electronic)
- SCSI (Small Computer System Interface)

Seriali:

- FireWire
- USB
- SDIO
- SATA (Serial ATA)
- SAS (Serially Attached SCSI)
- Fibre Channel (Fibre= Fiber+Wire)

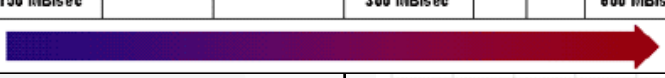
16

Dischi, interfacce, RAID

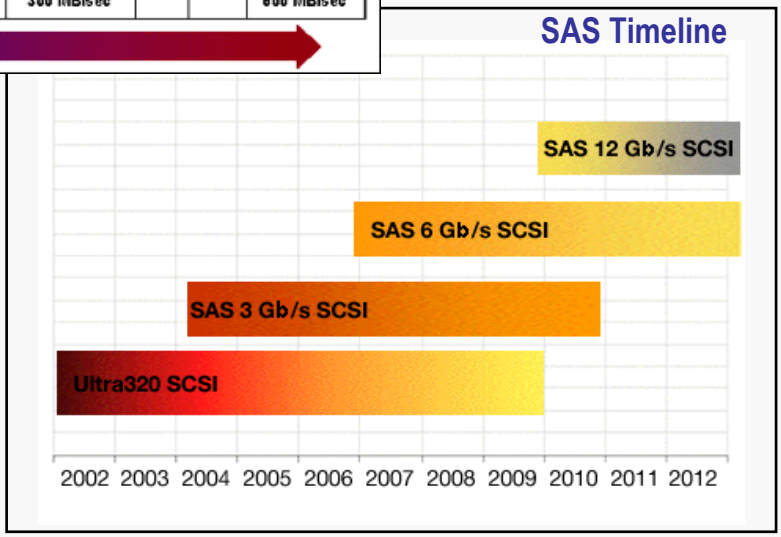
SATA e SAS: verso i 600 MB/sec

SATA Timeline

2001	2002	2003	2004	2005	2006	2007
1st Generation Specification	First SATA Products	Full support in Intel Chipset	2nd Generation Specification			3rd Generation Specification
150 MB/sec			300 MB/sec			600 MB/sec



SAS Timeline



17

Dischi, interfacce, RAID

Riassunto dello storage interface

	ATA	SATA	SCSI	SAS	FC
Number of devices	2	1	8 or 16	16K	16M
Maximum distance	18 ”	1 m	3-25 m	10 m	100+ km
Cable Type	Copper	Copper	Copper	Copper	Copper or fiber optic
Interface Type	Parallel	Serial	Parallel	Serial	Serial
Transfer speed (MB/s)	Up to 133	150,300	Up to 320	300	100,200,400
Hard drive rot. speed	Up to 7.2K	Up to 10K	10K, 15K	10K,15K	10K,15K

Alcune tendenze RAID

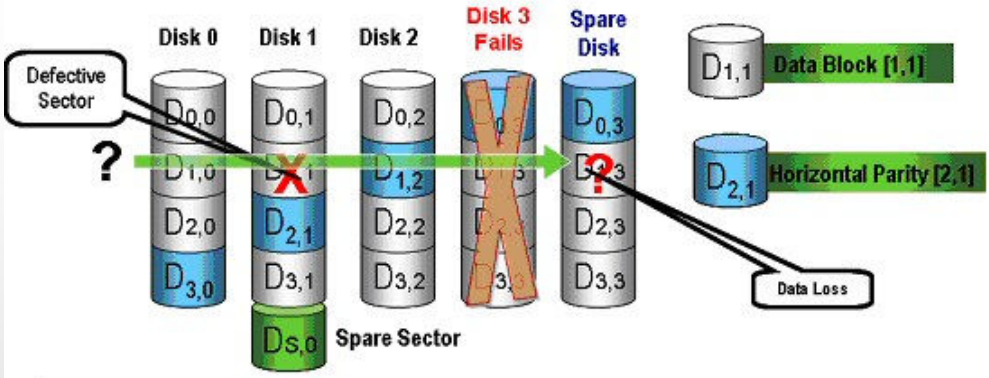
- E' in corso la campagna di migrazione verso il RAID-6. Il RAID-5 è resistente al fallimento di un disco. Ma dopo 2-3 anni di operazione i dischi si deteriorano e sono altamente probabili casi di fallimento completo di uno dei dischi in contemporanea a un accumulo di bad blocks non individuati su altri dischi (il RAID rebuild diventa in questo modo impossibile).

- Controller RAID-6 a basso costo diventano sempre più potenti. Modelli recenti (sia interni che esterni) sono in grado di erogare fino a 700-800 MB/sec (con dischi SAS).

- Cresce la popolarità dei sistemi RAID esterni con l'outlet IB-enabled.

Dischi, interfacce, RAID

Perchè RAID-6?



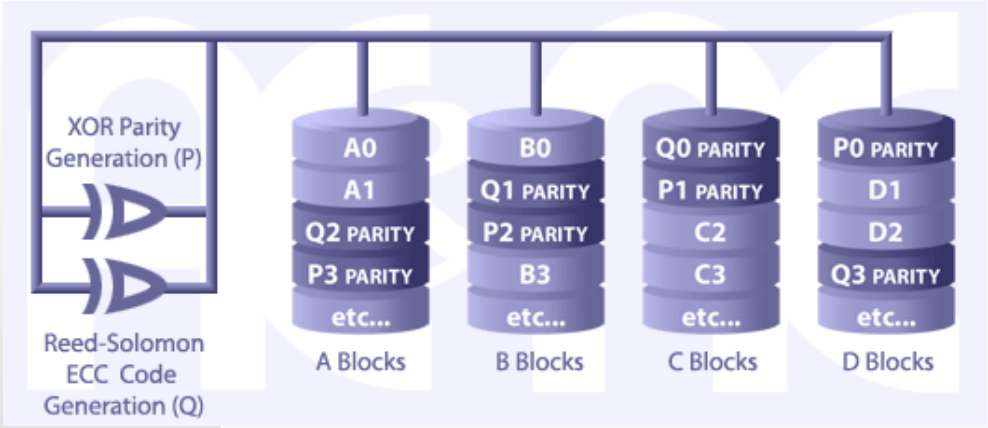
RAID-5 fails to handle accumulation of latent defects!

1. Latent defects accumulate *undetected* (even with disk auto-scrub)
2. Another disk in array eventually fails (sudden)
3. Rebuild to spare disk fails
4. Data loss – user data is lost!

20

Dischi, interfacce, RAID

RAID-6. Il problema con il RAID-5 appena illustrato può essere contrastato usando un nuovo livello di RAID che impiega il calcolo di due parity bit al posto di uno. Si possono utilizzare due XOR oppure uno XOR più un altro metodo di calcolo della parità:



21

Dischi, interfacce, RAID

Controller economici RAID-6. Tra i più recenti modelli interni da 16 dischi, buoni per le configurazioni "storage-in-a-box" si possono menzionare (prezzi E4 end-user):

- ARECA 1261ML-16P PCI-Express : 835 Euro + IVA
- AMCC(3Ware) 9650SE : 850 Euro + IVA
- Adaptec ICP 5165BR : 790 Euro + IVA

Per le configurazioni esterne da 16 dischi è raccomandabile:

- Infortrend A16F-G2430M2 SATA/FC : circa 4250 Euro + IVA (prezzo IFT)

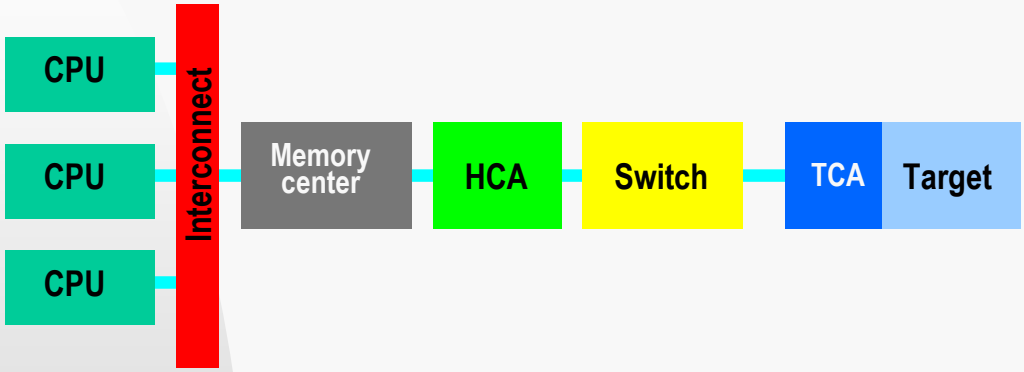
Abbiamo appena testato Areca e ICP, vanno a 320+ MB/sec con i RAID-6 da 8 drive SATA-II (large sequential writes su XFS). L'Infortrend in una configurazione equivalente eroga 400 MB/sec su un canale FC 4Gbit.

22

Dischi, interfacce, RAID

Sistemi RAID con l'outlet IB. Sono molto costosi ma offrono altissime prestazioni di picco spesso richieste per il calcolo HPC.

Protocolli:
SRP (SCSI RDMA Protocol: IEEE "Example of a storage protocol"),
iSER (iSCSI Extensions for RDMA, T10)



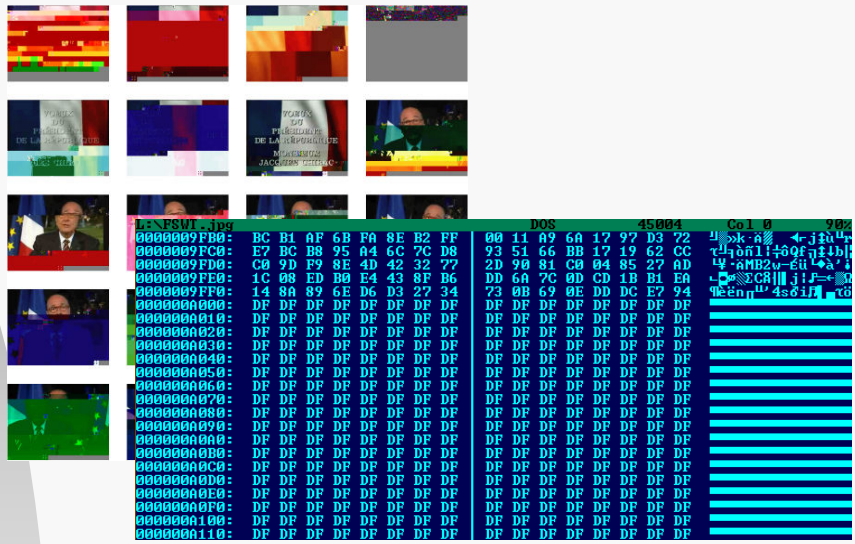
Alcuni sistemi RAID con l'interfaccia IB(SRP) già disponibili:

- DataDirectNetworks (DDN) S2A9500 (fino a 4 outlet IB) – 800+ MB/sec
- Mellanox MTD2000
- LSI Engenio 6498

In arrivo nel 2007: il primo sistema iSER (Voltaire/FalconStor)

23

Corruzioni silenziose



24

Corruzioni silenziose

Il 25 Aprile scorso Peter Kelemen (CERN) ha presentato i risultati di uno studio eseguito su gran parte dei loro sistemi storage. Lo studio aveva come finalità quella di stimare la probabilità di trovare dei dati registrati in precedenza (e in seguito intenzionalmente non modificati) nello stato diverso da quello originale.

Il problema delle corruzioni è ovviamente molto complesso e i fattori che possono contribuire sono tanti. Fra questi, citando Kelemen:

- Errori hardware (memory, CPU, disco, NIC)
- Data Transfer Noise (UTP, SATA, FC, Wireless)
- Bachi nel firmware (RAID controller, disco, NIC)
- Bachi software, kernel, VM, block layer
- Errori File System

Un'applicazione di prova appositamente sviluppata è stata fatta girare su 3500 nodi per un periodo abbastanza lungo. Il traffico dati complessivo è stato di circa 41PB. Sono stati scoperti errori su 170 nodi.

Corruzioni silenziose

Analizzando i dati raccolti, è stato scoperto che:

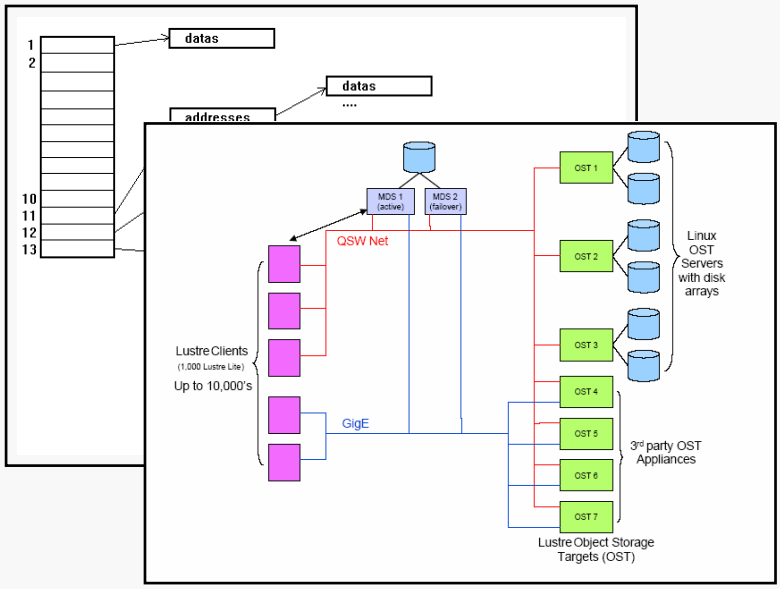
- La frequenza degli errori è compatibile con l'expected Bit Error Rate delle varie componenti
- Gli errori sono distribuiti uniformemente nel tempo
- Sono ben individuabili 3 tipologie distinte di errori:
  1) Errori persistenti di singolo o doppio flip dei bit nel byte riconducibili a problemi di memoria accertati
  2) Errori transienti di "incursioni di small chunk di dati randomici"
  3) Errori persistenti di apparizione di chunk multipli da 64KB, molto correlati a I/O command timeout (la stragrande maggioranza di errori è di questo tipo)

Stanno lavorando sul problema. Al momento le conclusioni tratte sono le seguenti:

- Le corruzioni silenziose sono "un fatto di vita"
- Non sarà mai possibile eliminare tutti gli errori completamente ..
- .. però bisogna essere in grado di diagnosticarli
- Le spese necessarie per garantire l'integrità dei dati sono molto elevate: copie multipli (+disco), algoritmi/procedure di correzione (+CPU)

IMHO, se gli errori di tipo 1 e 3 sono riconducibili a eventi noti occorrerebbe allora vedere se sia possibile indagare su questi eventi (girare regolarmente i controlli di memoria, scoprire perchè e quando succedono gli I/O command timeout e se possibile fare qualcosa a tal proposito)...

Notizie FS (locali e distribuiti)



Notizie FS

Ext4: meno limiti, più performance per l'ExtX

- Introdotto nell'Ottobre 2006 (kernel 2.6.19)
- Lo sviluppo è stato motivato dalle limitazioni di Ext3:
  - 16 TB file system limit (numero di blocchi a 32 bit)
  - Timestamp troppo impreciso (un secondo)
  - Limite basso per il numero di entry per subdirectory (32768)
  - Prestazioni basse per i file grandi
- 48 bit riservati per numero di blocchi (limite di 1 EB per file system), 64 bit per i metadati ove necessario (nel SB, journal) – limite di 16 TB per file
- Migrazione dalle indirect block maps alla logica degli extents:
  - Le indirect block maps diventano poco efficienti per file grandi (un extra block read e seek ogni 1024 blocchi)
  - Un extent descrive invece un gruppo di blocchi contigui (un modo efficiente di rappresentare file grandi)
  - Molto meno I/O per i metadati, meno lavoro per la CPU

28

Notizie FS

- Lavori Ext4 in corso (patch disponibili):
  - Allocazione simultanea per blocchi multipli contigui
  - Delayed block allocation (sospensione di allocazione fino a write back)
  - Persistent file preallocation (preallocare senza inizializzare)
  - Defragmentazione online (nel kernel)
  - Nanosecond timestamp (serve per NFSv4)
  - Numero illimitato di subdirectory
  - Metadata checksumming

- Come migrare a Ext4:

mount –t ext4dev	(vale anche per i file system ext3 preesistenti)
mount –t ext4dev –o extents	(abilita extents, rimane compatibile con ext3 fino alla prima operazione di scrittura)

29

Notizie FS

ZFS: sta diventando sempre più utile

- Introdotto dalla Sun nel 2004. Ora è usato con successo per dCache storage pools e potrebbe essere preso in considerazione per i file server NFS per grandi file sulle macchine commodity x86 con Solaris OS.

- Perchè ZFS? Innanzitutto per l'elevata integrità dei dati in esso archiviati:
  - TRANSACTIONAL (NO VFS LAYER) – integrità dati end-to-end
  - Flexible storage pools
  - Software RAID (RAID-Z) in combinazione con transactions è molto affidabile
  - Fsck: non serve

- Alcuni dettagli:
  - Solo 2 comandi gestionali da conoscere – "zpool" e "zfs"
  - Dimensioni di file e file system enormi (128 bit base)
  - Snapshots, ACL ecc.
  - Da verificare: prestazioni per file piccoli

- Deployed base nell'HEP: 800 TB IN2P3 (dCache, xrootd), 200 TB DESY (dCache)

30

Notizie FS

Storage-in-a-box ad alta densità basato su ZFS ~ 2 USD/GB

- Popolarissimo (DESY, SLAC, IN2P3 e altri) : Sun Thumper (Sun Fire X4500)

- 24 TB in 4U
- 2 dual core AMD 285 (2.6 Ghz)
- RAM: max 16 GB
- 4 Gigabit Ethernet outlet
- RAID-Z (software 0,1,5)
- 1.5 -1.8 KW

31

Notizie FS

AFS: concluso il progetto AFS/OSD

- Il progetto è stato avviato nella primavera del 2005. Puro R&D, il fine era di aumentare la scalabilità e possibilmente la performance utilizzando le tecnologie OSD (Object Shared Device, T10).

- Performance: abbiamo scoperto un difetto nell'implementazione dello stack del principale protocollo di trasmissione dati AFS (Rx). In presenza di certe race conditions possono verificarsi delle ritrasmissioni. Le prestazioni di picco dipendono dalla situazione LAN, dal router e dalla potenza di picco della CPU usata sulle macchine server. Abbiamo stimato che per risolvere il problema (che abbiamo sottoposto a un gruppo OpenAFS) occorrono circa 2 anni uomo; abbiamo quindi sospeso i lavori in questa direzione. Allo stato attuale la peak performance ottenibile oscilla tra i 50 e 70 MB/sec.

- Scalabilità: in questo campo sono stati raggiunti tutti gli obiettivi. Ora è diventato possibile:
  - Stendere un volume R/W su più OSD
  - Avere più copie R/W dello stesso file
  - Mantenere la vecchia organizzazione dati/client/server e sfruttare gli OSD ove possa essere utile

- Chiusura del progetto:
  - La presentazione finale è stata fatta da H.Reuter il 25 Aprile 2007
  - Il codice è disponibile e sarà tra breve sottomesso a OpenAFS

32

Notizie FS

Lustre: è appena uscita la versione 1.6.0

- Novità principali:
  - La gestionabilità è decisamente migliorata: finalmente ci sono mkfs e mount “cristiani”
  - E’ diventato possibile aggiungere gli OST a caldo
  - Non è più necessario sostituire il kernel a livello di client
  - Ottimo supporto RDMA, fino a 1 Gbyte/sec peak perfomance su IB
  - E’ possibile configurare i dischi back-end OST nella variante ridondata (primary/backup)
- Implementazione elementi ILM:
  - Sono in corso le attività di sviluppo dei feature di true HSM (con HPSS e altri)
  - Il rilascio è previsto per 2008

GPFS: verso il supporto RDMA nativo

- Il rilascio del supporto RDMA nativo è promesso per l’estate 2007
- Studio e implementazione elementi ILM: in corso (con HPSS)

pNFS e NFSv4/RDMA: da tenere sott’occhio

- Tante attività di sviluppo, primi benchmark interessanti
- Non ha raggiunto ancora la production quality

33

Seguono i lucidi dell'HEPiX FSWG Progress Report

34


HEPiX FSWG Progress Report
Andrei Maslennikov
April 2007 - Hamburg

Summary

- Raison d'être
- Members
- General Plan
- Workflow February - April 2007
- Immediate goals for May - September 2007
- Discussion

Raison d'être

- The group was commissioned by IHEPCCC in the end of 2006
- Officially supported by the HEP IT managers
- The goal is to review the available file system solutions and storage access methods, and to divulge the know-how among HEP organizations and beyond

- Timescale : Feb 2007 – April 2008
- Milestones: 2 progress reports (Spring 2007, Fall 2007),
  1 final report (Spring 2008)

37

Members

- Currently we have 20 people on the list, but only these 15 appeared in the meetings/conf. calls and did something since the group had started:

BNL	R.Petkus
CASPUR	A.Maslennikov (Chair), M.Calori (Web Master)
CEA	J-C.Lafoucriere
CERN	B.Panzer-Steindel
DESY	M.Gasthuber, P.van der Reest
FZK	J.van Wezel, S.Meier
IN2P3	L.Tortay
INFN	V. Sapunenko
LAL	M.Jouvin
NERSC/LBL	C.Whitney
RZG	H.Reuter
U.Edinburgh	G.A.Cowan

- These very people maintain contacts with several other important labs like LLNL, SLAC, JLAB, DKRZ, PNL and others.

38

General Work Plan

- The work plan for the group was discussed and agreed upon during the first two meetings. Accent will be made on shared / distributed file systems.
- We start with an **Assessment** of the existing file system / data access solutions; at this stage we will be trying to classify the storage use cases
- Next, in the course of the **Analysis** stage we will try get a better idea of the requirements for each of the classes defined during the previous stage
- This will be followed by the selection of the viable **Candidate Solutions** for each of the storage classes, followed by a possible **Evaluation** of some of them on the common hardware
- Then the **Final Report** with conclusions and practical recommendations will be due, by the Spring 2008 HEPiX meeting

39

Assessment

- Understand how storage is accessed and used in our organizations and beyond. Classify storage solutions in function of typology of applications that make use of the data areas. Possible storage classes may include:
  - Areas for home directories
  - Areas for software repositories
  - Areas for large data archival and processing
  - Areas for data analysis
  - Areas for databases
  - Areas for engineering applications
  - Areas for HPC applications
  - Temporary (scratch) areas
  - Other areas

Analysis

- Try to evidentiate specific requirements for the previously defined storage classes. For each of the classes, mention:
  - Typical access patterns (TFA/non, streaming, random, block sizes etc)
  - Minimal acceptable access performance (peak and sustained)
  - Typical sizes of areas and files stored therein
  - Should these areas be shared among many machines
  - Should these areas be shared geographically
  - Scalability requirements
  - Redundancy requirements
  - GRID compatibility
  - HSM requirements
  - Typical benchmarks relevant for this class

41

Candidate Solutions

- Prepare a list of known practicable FS / data access solutions for each of the storage classes. This list could include Castor, dCache, AFS, GPFS, Lustre, SRM, NFS, Panasas, Teragrid, Ibrix, ZFS, XFS, StorNext, GFS, PVFS, PNFS, CXFS and others. For each of the solutions, mention:
  - Its specific features
  - Software costs (purchase and maintenance), if there are any
  - Hardware costs
  - Hardware/software dependencies and limitations
  - Size of the installed base
  - I/O rates typically observed in our orgs and beyond
  - Complexity / ease of use, FTEs needed to keep it alive

Workflow 08/02/07-12/04/07

- Set up the mailing list and web site
- Started with the Assessment step
- Held 6 regular phone and one out of band face-to-face meetings
- Established contacts with the organizations that do not participate in our working group directly

Assessment progress

- Prepared an online questionnaire on deployed file stores
- Selected 21 important sites to be covered: Tier-0, all Tier-1 plus several large labs/orgs like CEA, LLNL, DKRZ
- All selected sites were invited to fill the questionnaire for their most relevant file store solutions; at least two areas had to be covered: home directories and the largest available shared filestore

Sites under assessment

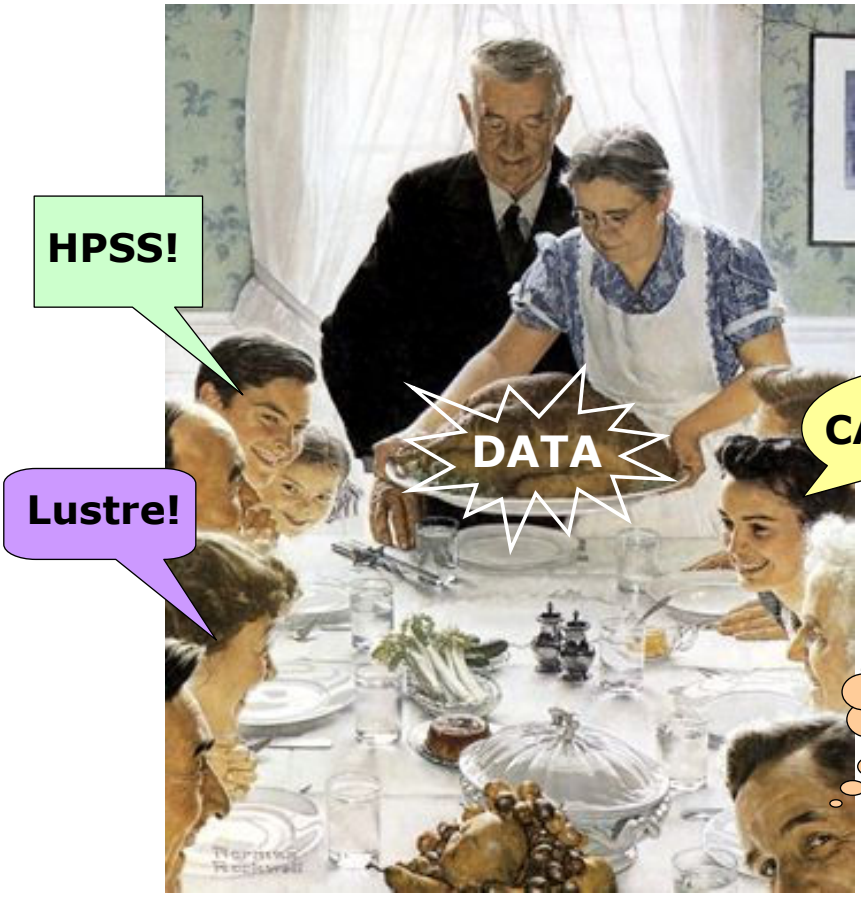
- ASGC
- BNL
- CC-IN2P3
- CEA
- CERN
- CNAF
- DAPNIA
- DESY
- DKRZ
- FNAL
- FZK
- JLAB
- LLNL
- NERSC
- Netherlands LHC
- NDGF
- PIC
- PNL
- RAL
- RZG
- SLAC
- TRIUMF

- - Collected / being verified
- - Being collected
- - NO INFO / NO CONTACT

45

Initial observations

- Not a big surprise, but:
  - All sites are unequal
  - Some sites are more unequal than others
  - Most of the sites collaborated…
  - .. but not all of them were able to provide the needed information (in particular, some questions on performance were not answered)
- The big picture looks a bit chaotic, the reasons to choose this or that storage access platform are often not clear
- The good side of the medal is that having a variety of solutions one may compare them and pronounce on all of them.

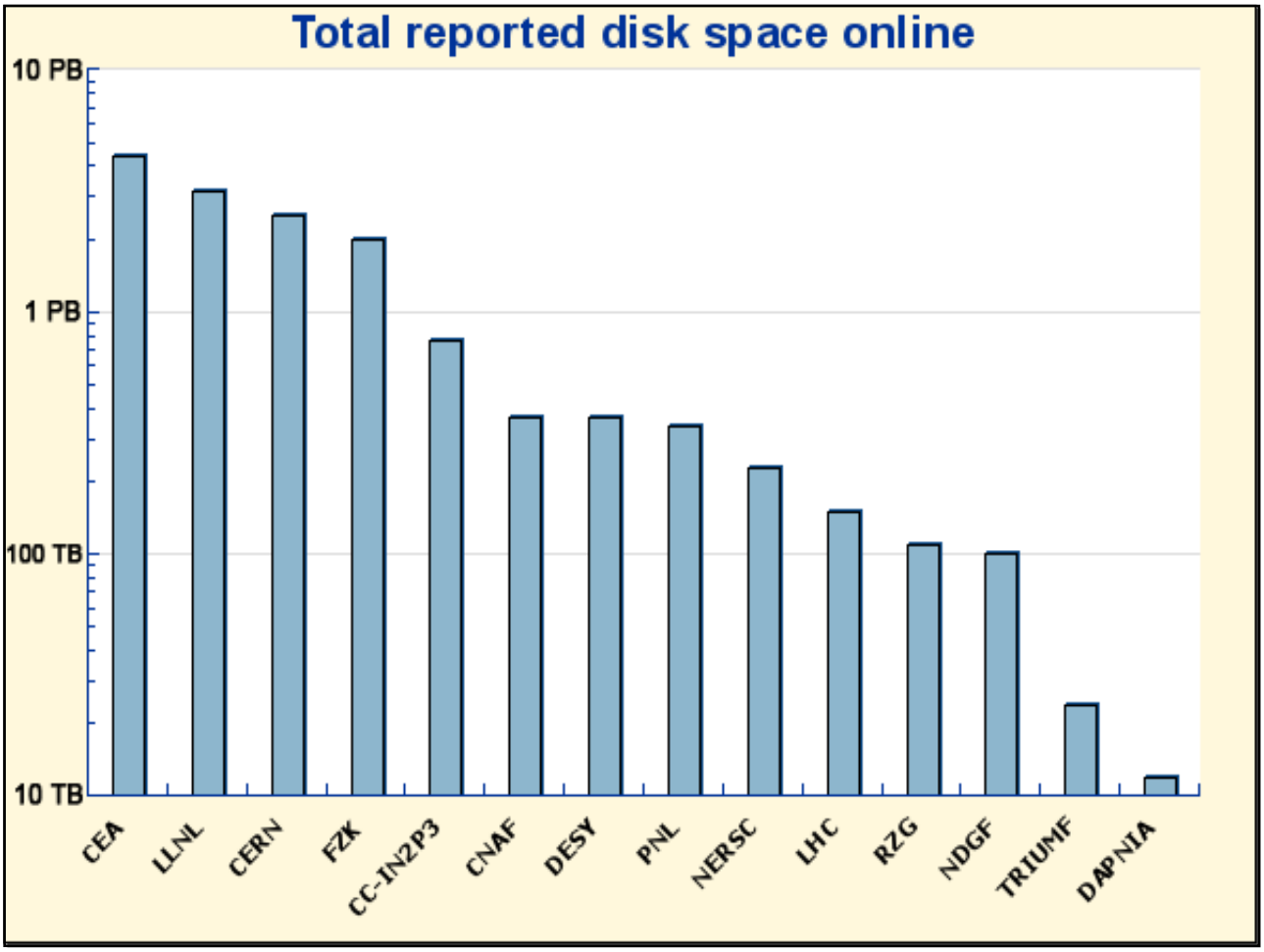


47

Some comments and first bulk numbers

- The data collected are yet to be verified! Some of the numbers provided by the local “info collectors” appear to be unprecise.
- Moreover, in several cases some fields of the questionnaire were interpreted in different ways by different info collectors. We hence scheduled an effort to clean this up (“normalize”), and to see if the questions should be made in a better form.
- So far we were only able to make a pair of intermediate plots on the basis of data collected over 14 sites out of planned 21, but already these partial infos could tell us something. We only looked at the online disk areas. The slow tier (tape backend) has to be studied seprately, and we still miss plenty of data.
- The total area size online reported is large but may not be called very impressive: 13.7 PB over all sites including the non-HEP organizations (compare with the planned 12-14 PB/year for LHC production).

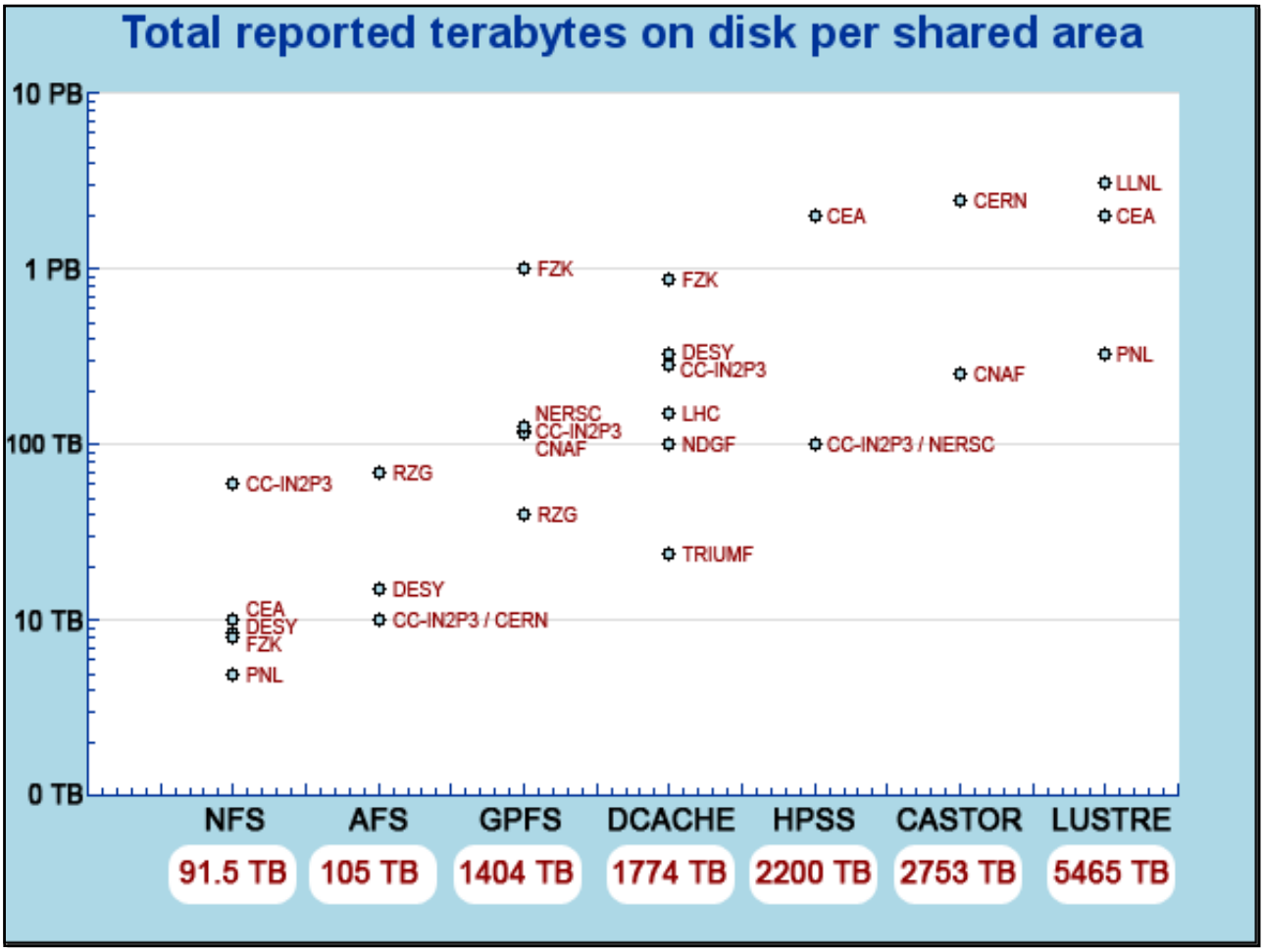
48


Total reported disk space online
10 PB
1 PB
100 TB
10 TB
CEA
LLNL
CERN
FZK
CC-IN2P3
CNAF
DESY
PNL
NERSC
LHC
RZG
NDGF
TRIUMF
DAPNIA

File systems / data access solutions in use

- Please note that we are still *very* far away from any conclusions!
- However, here are some facts and thoughts:
  - The large initial list of candidate solutions may probably be reduced to just 7 names: Lustre, GPFS, HPSS, CASTOR, dCache, AFS and NFS
  - AFS and NFS are mostly used for home directories and software repositories and remain very popular
  - Solutions with the HSM function (HPSS, CASTOR and dCache) have similar deployed base in petabytes, and probably have to be compared
  - GPFS and Lustre dominate in the field of distributed file systems and deserve to be compared
  - Lustre has the largest reported installed base (5.5 PB), but not a single HEP organization had ever deployed it !
  - dCache is present in many HEP sites, however CASTOR alone stores more data than all reported dCache areas (NB: we miss data from FNAL)

50


Total reported terabytes on disk per shared area
10 PB
1 PB
100 TB
10 TB
0 TB
CC-IN2P3
CEA
DESY
FZK
PNL
RZG
DESY
CC-IN2P3 / CERN
FZK
NERSC
CC-IN2P3
CNAF
RZG
FZK
DESY
CC-IN2P3
LHC
NDGF
TRIUMF
CEA
CC-IN2P3 / NERSC
CERN
CNAF
LLNL
CEA
PNL
NFS
AFS
GPFS
DCACHE
HPSS
CASTOR
LUSTRE
91.5 TB
105 TB
1404 TB
1774 TB
2200 TB
2753 TB
5465 TB

Tentative plan until September 2007

- Continue with the data collection and analysis
  - Complete the questionnaire by the Fall 2007
  - Report during the meeting at St Louis
- Reduce the list of solutions to 7 names and create three mini task forces:
  - On home directories / software repositories: AFS, NFS, GPFS(?)
  - On data access solutions with the tape backend: CASTOR, dCache, HPSS
  - On scalable high performance distributed file systems: Lustre, GPFS
- Each of the task forces will have a goal to prepare an exhaustive collection of documentation on the corresponding solutions, describe best practices, provide deployment advice, cost estimates and performance benchmarks
  - All task forces will have to present an interim progress report during the St Louis meeting

52

Some input for discussion

- This workgroup is open to all sites (HEP- and non-) interested in the storage issues. We appreciate any feedback and would welcome any new active members
- We appeal to FNAL to join us actively (or at least to provide their data on storage, otherwise our report will not be complete)
- Our web site (http://hepix.caspur.it/storage) is open to all universities, research labs and organizations. The access to it is protected by a symbolic password which was widely circulated among HEPiX members and may at any time be obtained via mail. Just send your request to *monica.calori at caspur.it*

53