Mass Storage System for Disk and Tape resources at the Tier1.

Ricci Pier Paolo *et al.*, on behalf of INFN TIER1 Storage

pierpaolo.ricci@cnaf.infn.it

ACAT 2008

November 3-7, 2008

Erice

Summary

- Tier1 Disk and Tape resources
- Castor status
- Disk SAN and GPFS
- TSM (tape backend for GPFS)
- GPFS and TSM first results

Tier1 Disk and Tape resources

- Here is what we have in production:

Disk (SAN): ~1250 TB RAW (ATA RAID-5)

9 Infortrends A16F-R1211-M2 50TB

2 SUN STK Bladestore 80TB

4 IBM FastT900 (DS 4500) 160TB

5 SUN STK FLX600 290TB

3 DELL EMC CX-380 670TB

Installation of additional 8 DELL EMC 1600TB in progress NEXT MONTH => **2.5 PBYTE**

Tier1 Disk and Tape resources

Tape: 2 tape robot libraries in production

1 SUN STK L5500 partitioned in 2000 slots LTO-2 (200GB) and 3500 slots 9940B (200GB)

6 LTO-2 Drives (20-30 MB/s each)

10 9940B Drives (25-30 MB/s each)

1 Pbyte Capacity

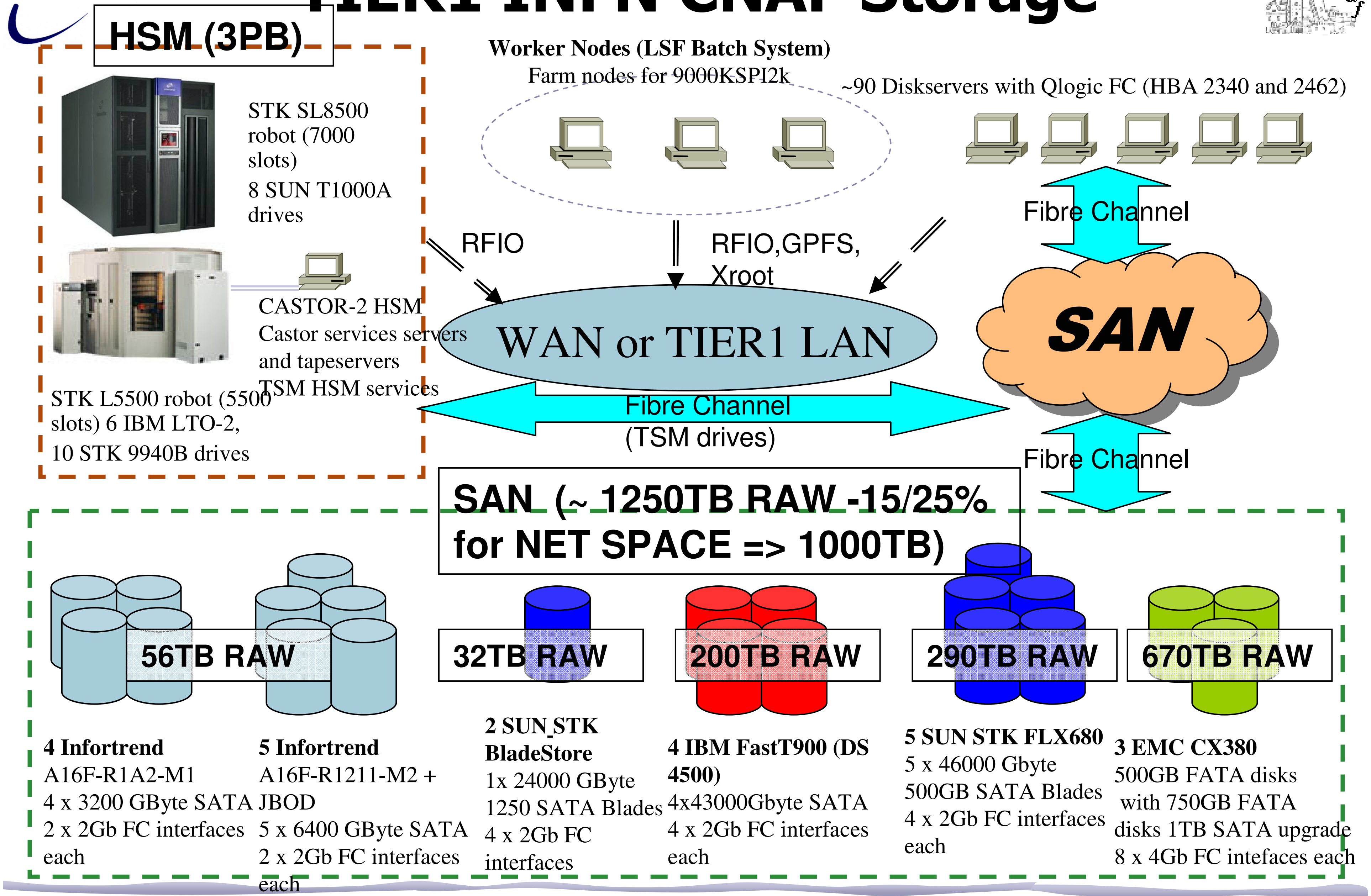
1 SUN SL8500 with 7000 slots T1000 slot (4000 tapes)

8 T1000A Drives (500GB/tape capacity and 110 MB/s bandwidth) in production

2 Pbyte Actual Capacity

UPGRADE to 10000 slots and 20 T1000B Drives (1TB/tape capacity) at end 2008 => 10 Pbyte capacity

TIER1 INFN CNAF Storage



HSM (3PB)

STK SL8500 robot (7000 slots)

8 SUN T1000A drives

CASTOR-2 HSM Castor services servers and tapeservers

TSM HSM services

STK L5500 robot (5500 slots) 6 IBM LTO-2, 10 STK 9940B drives

Worker Nodes (LSF Batch System)

Farm nodes for 9000KSPI2k

~90 Diskservers with Qlogic FC (HBA 2340 and 2462)

SAN (~ 1250TB RAW -15/25% for NET SPACE => 1000TB)

4 Infortrend
A16F-R1A2-M1
4 x 3200 GByte SATA
2 x 2Gb FC interfaces each

5 Infortrend
A16F-R1211-M2 + JBOD
5 x 6400 GByte SATA
2 x 2Gb FC interfaces each

2 SUN STK BladeStore
1x 24000 GByte
1250 SATA Blades
4 x 2Gb FC interfaces

4 IBM FastT900 (DS 4500)
4x43000Gbyte SATA
4 x 2Gb FC interfaces each

5 SUN STK FLX680
5 x 46000 Gbyte
500GB SATA Blades
4 x 2Gb FC interfaces each

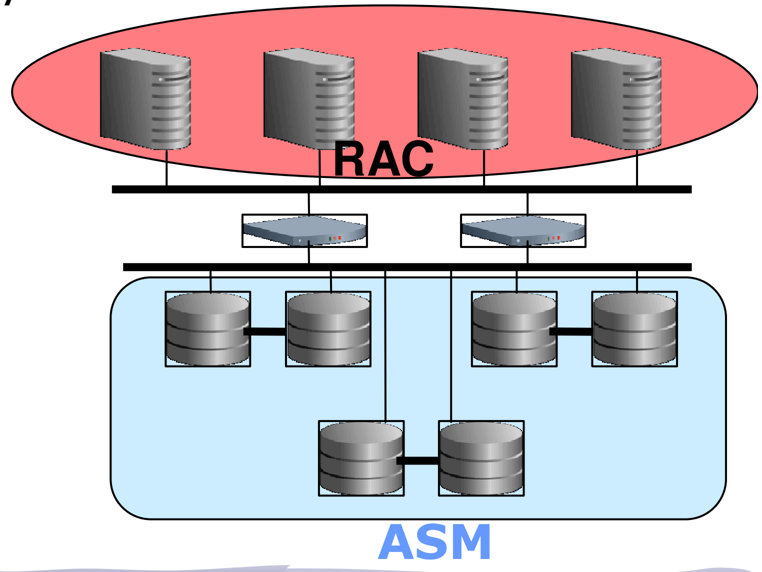
3 EMC CX380
500GB FATA disks with 750GB FATA disks 1TB SATA upgrade
8 x 4Gb FC intefaces each

Oracle Database Service

- Main goals: high availability, scalability, reliability
- Achieved through a modular architecture based on the following building blocks:
  - Oracle ASM volume manager for storage management implementation of redundancy and striping in an Oracle oriented way
  - Oracle Real Application Cluster (RAC) the database is shared across several nodes with failover and load balancing capabilities (Castor with 5 instances, LCG File Catalog Atlas LHCB, Lemon, SRM)
  - Oracle Streams geographical data redundancy for LHCB conditions database

- 32 server, 24 of them configured in 12 cluster
- 30 database instances
- Storage: 5TB FC Array dedicated (20TB raw) UPGRADE TO 40TB raw (installing now...)
- **Availability rate: 98,7% in 2007**

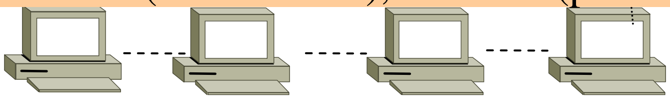
Availability (%) = Uptime/(Uptime + Target Downtime + Agent Downtime)



CASTOR 2.1.7-17 deployment

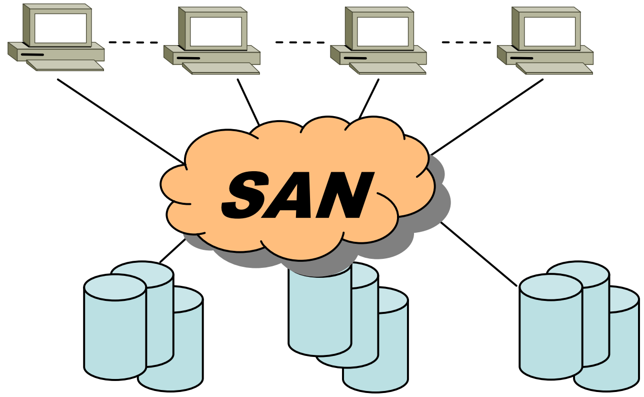
- Core services are on machines with SCSI disks, hardware RAID1, redundant power supplies SLC4 32 bit
- tape servers and disk servers have lower level hardware, like WNs

- STK L5500 silos (5500 slots, 200GB cartridges, capacity ~1.1 PB) + SL8000 silos (7000 slots, 500GB/1TB cartridges, actual capacity ~2 PB)
- 24 tape drives, 3 Oracle databases (DLF, Stager, Nameserver) on ORACLE Real Application Cluster
- LSF plug-in for scheduling
- SRM v2 (2 front-ends), SRM v1 (phasing out)



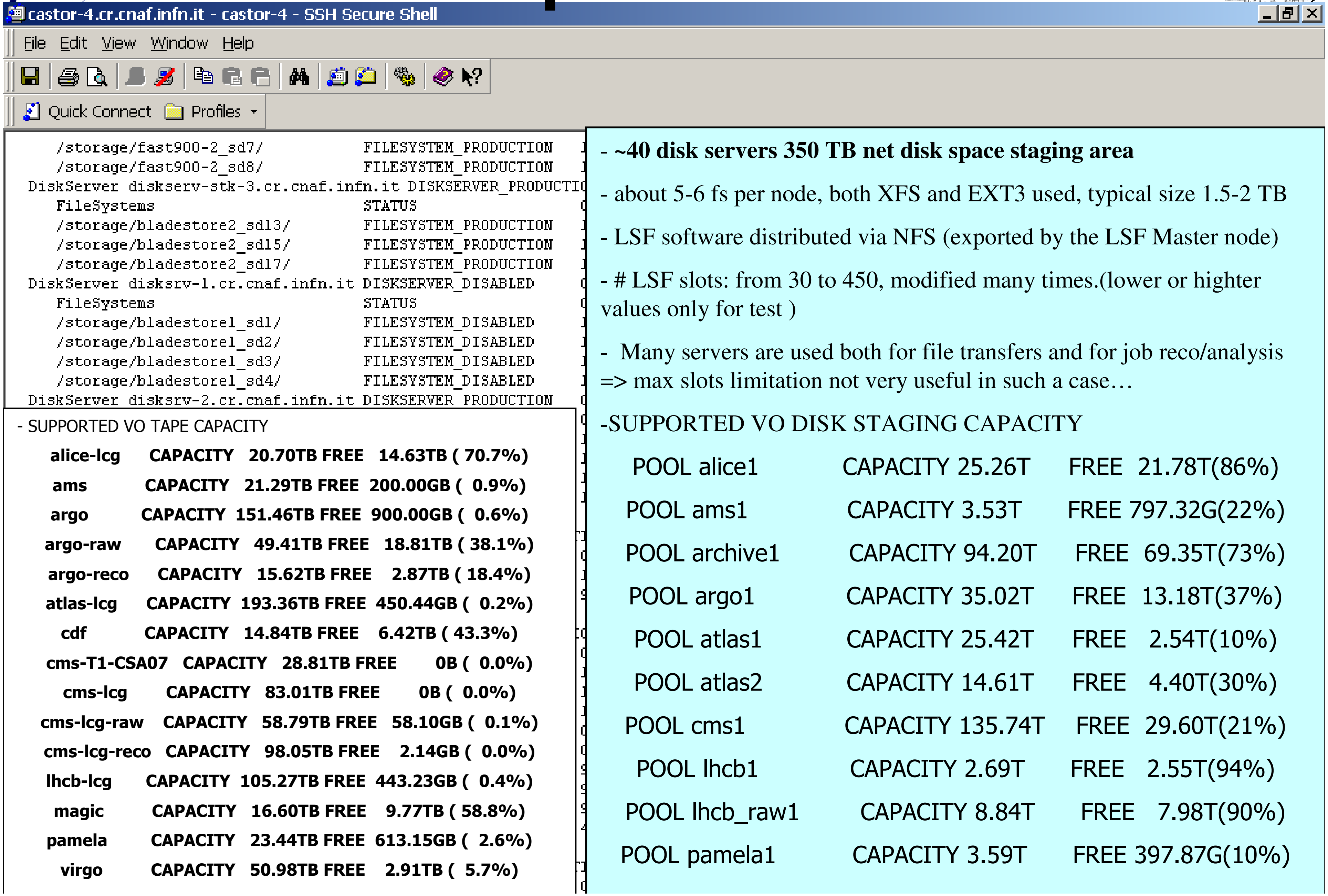
15 tape servers

~ 40 CASTOR disk servers attached to a SAN full redundancy FC 2Gb/s or 4Gb/s connections (dual controller HW and Qlogic SANsurfer Path Failover SW or Vendor Specific Software)



STK FlexLine 600...

CASTOR setup



```
    /storage/fast900-2_sd7/           FILESYSTEM_PRODUCTION
    /storage/fast900-2_sd8/           FILESYSTEM_PRODUCTION
 DiskServer diskserv-stk-3.cr.cnaf.infn.it DISKSERVER_PRODUCTIO
    FileSystems                       STATUS
    /storage/bladestore2_sdl3/        FILESYSTEM_PRODUCTION
    /storage/bladestore2_sdl5/        FILESYSTEM_PRODUCTION
    /storage/bladestore2_sdl7/        FILESYSTEM_PRODUCTION
 DiskServer disksrv-1.cr.cnaf.infn.it DISKSERVER_DISABLED
    FileSystems                       STATUS
    /storage/bladestore1_sdl/         FILESYSTEM_DISABLED
    /storage/bladestore1_sd2/         FILESYSTEM_DISABLED
    /storage/bladestore1_sd3/         FILESYSTEM_DISABLED
    /storage/bladestore1_sd4/         FILESYSTEM_DISABLED
 DiskServer disksrv-2.cr.cnaf.infn.it DISKSERVER_PRODUCTION
```

- SUPPORTED VO TAPE CAPACITY

alice-lcg	**CAPACITY**	**20.70TB**	**FREE**	**14.63TB**	**(70.7%)**
ams	**CAPACITY**	**21.29TB**	**FREE**	**200.00GB**	**(0.9%)**
argo	**CAPACITY**	**151.46TB**	**FREE**	**900.00GB**	**(0.6%)**
argo-raw	**CAPACITY**	**49.41TB**	**FREE**	**18.81TB**	**(38.1%)**
argo-reco	**CAPACITY**	**15.62TB**	**FREE**	**2.87TB**	**(18.4%)**
atlas-lcg	**CAPACITY**	**193.36TB**	**FREE**	**450.44GB**	**(0.2%)**
cdf	**CAPACITY**	**14.84TB**	**FREE**	**6.42TB**	**(43.3%)**
cms-T1-CSA07	**CAPACITY**	**28.81TB**	**FREE**	**0B**	**(0.0%)**
cms-lcg	**CAPACITY**	**83.01TB**	**FREE**	**0B**	**(0.0%)**
cms-lcg-raw	**CAPACITY**	**58.79TB**	**FREE**	**58.10GB**	**(0.1%)**
cms-lcg-reco	**CAPACITY**	**98.05TB**	**FREE**	**2.14GB**	**(0.0%)**
lhcb-lcg	**CAPACITY**	**105.27TB**	**FREE**	**443.23GB**	**(0.4%)**
magic	**CAPACITY**	**16.60TB**	**FREE**	**9.77TB**	**(58.8%)**
pamela	**CAPACITY**	**23.44TB**	**FREE**	**613.15GB**	**(2.6%)**
virgo	**CAPACITY**	**50.98TB**	**FREE**	**2.91TB**	**(5.7%)**

- **~40 disk servers 350 TB net disk space staging area**

- about 5-6 fs per node, both XFS and EXT3 used, typical size 1.5-2 TB

- LSF software distributed via NFS (exported by the LSF Master node)

- # LSF slots: from 30 to 450, modified many times.(lower or highter values only for test)

- Many servers are used both for file transfers and for job reco/analysis => max slots limitation not very useful in such a case…

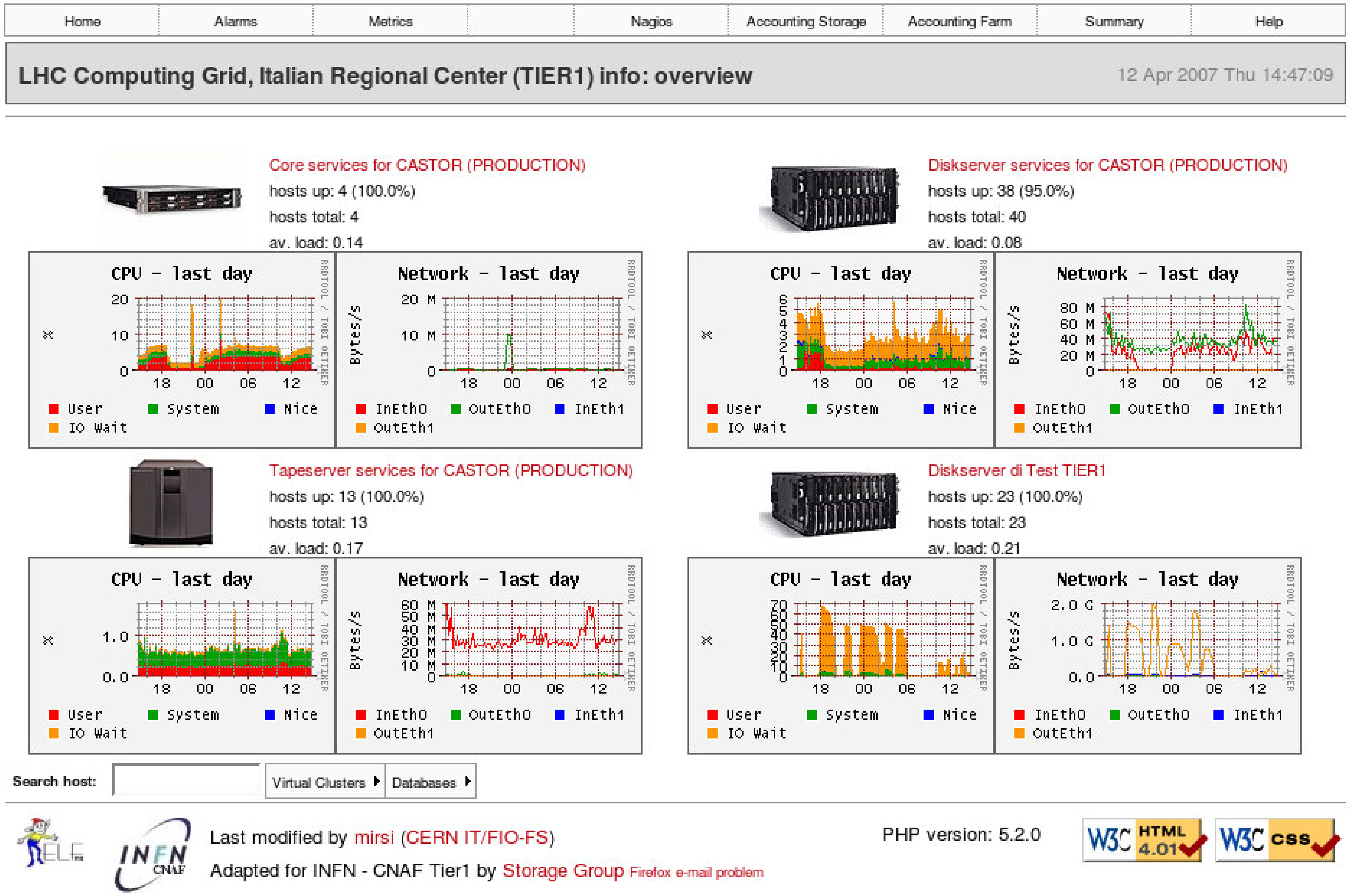
-SUPPORTED VO DISK STAGING CAPACITY

POOL alice1	CAPACITY 25.26T	FREE 21.78T(86%)
POOL ams1	CAPACITY 3.53T	FREE 797.32G(22%)
POOL archive1	CAPACITY 94.20T	FREE 69.35T(73%)
POOL argo1	CAPACITY 35.02T	FREE 13.18T(37%)
POOL atlas1	CAPACITY 25.42T	FREE 2.54T(10%)
POOL atlas2	CAPACITY 14.61T	FREE 4.40T(30%)
POOL cms1	CAPACITY 135.74T	FREE 29.60T(21%)
POOL lhcb1	CAPACITY 2.69T	FREE 2.55T(94%)
POOL lhcb_raw1	CAPACITY 8.84T	FREE 7.98T(90%)
POOL pamela1	CAPACITY 3.59T	FREE 397.87G(10%)

Castor Monitoring (Lemon)

- Lemon is in production as a Monitoring Tool
- Lemon is the CERN suggested monitoring tool, strong integration with Castor v.2
- Oracle10 on Real Application Cluster as database backend

Lemon Monitoring



LHC Computing Grid, Italian Regional Center (TIER1) info: overview

12 Apr 2007 Thu 14:47:09

Core services for CASTOR (PRODUCTION)
hosts up: 4 (100.0%)
hosts total: 4
av. load: 0.14

Diskserver services for CASTOR (PRODUCTION)
hosts up: 38 (95.0%)
hosts total: 40
av. load: 0.08

Tapeserver services for CASTOR (PRODUCTION)
hosts up: 13 (100.0%)
hosts total: 13
av. load: 0.17

Diskserver di Test TIER1
hosts up: 23 (100.0%)
hosts total: 23
av. load: 0.21

Last modified by mirsi (CERN IT/FIO-FS)
Adapted for INFN - CNAF Tier1 by Storage Group Firefox e-mail problem

PHP version: 5.2.0

STORAGE AREA NETWORK

All Disk Hardware at our Tier1 is on Storage Area Network.

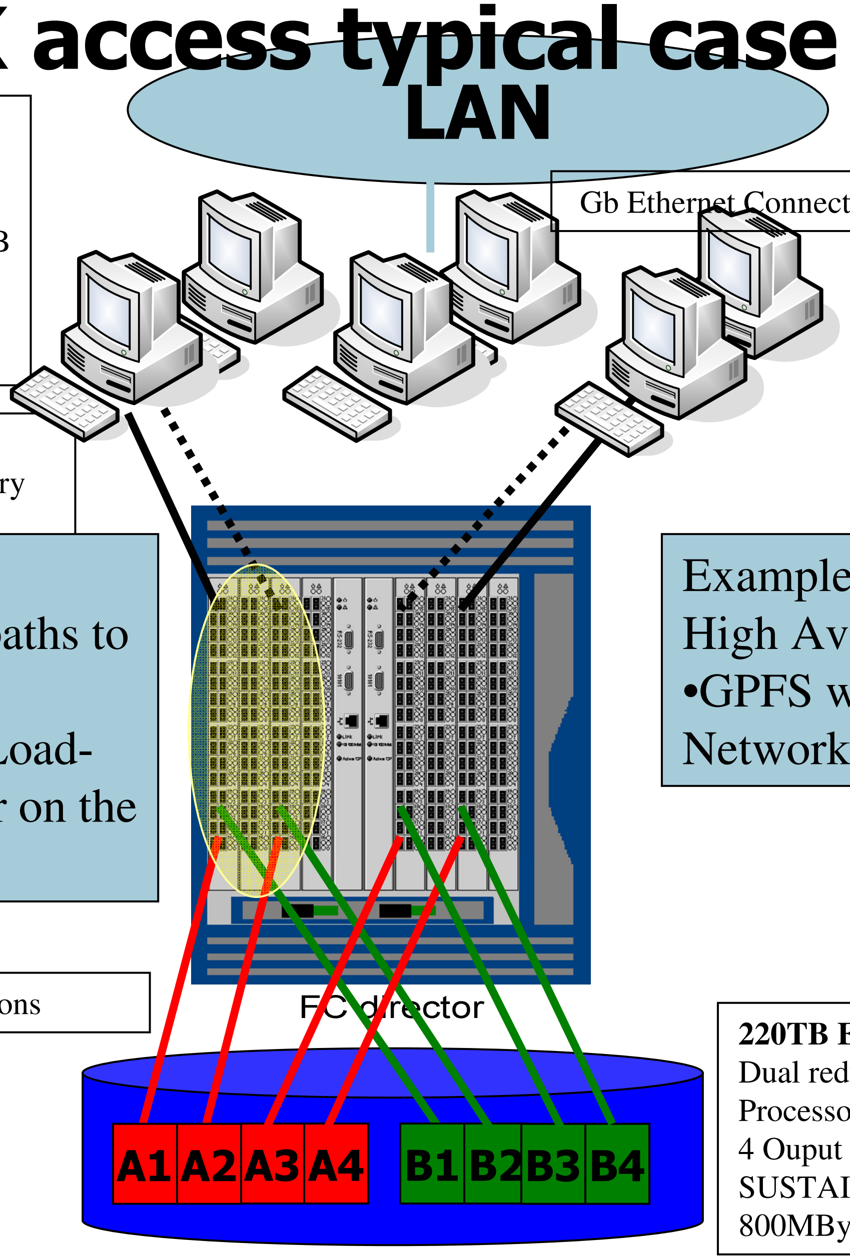
SAN give some good advantages:

- diskservers could implement a No Single Point of Failure system where every component of the storage system is rendundant (storage array controllers, SAN switches, and server HBA). If software supports it, a cluster approach is possible
- The SAN give the best flexibility, we can dinamically assign new volumes or disk storage arrays to diskservers
- Monitoring tool on SAN could help to monitor i/o bandwidth on devices
- LAN free systems for archiving and backup purpose to the tape facilities is possible

DISK access typical case (NSPF)

12 Diskserver Dell 1950
Dual Core Biprocessors
2 x 1.6Ghz 4MB L2 Cache,
4 GByte RAM, 1066 MHz FSB
SL 3.0 or 4.0 OS, Hardware
Raid1 on system disks and
redundant power supply

2 x 4Gb Qlogic 2460 FC redundand connections every Diskserver



LUN0
LUN1
...

LUN0 => /dev/sda
LUN1 => /dev/sdb
...

SAN ZONING:
Each diskserver => 4 paths to the storage
•EMC PowerPath for Load-Balancing and Failover on the 4 paths

Example of Application High Avaliability:
•GPFS with configuration Network Shared Disk

4Gb FC connections

2 Storage Processor (A e B)

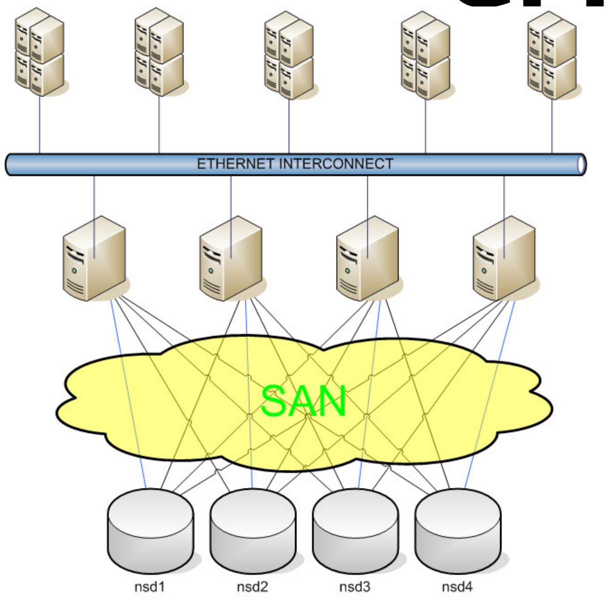
220TB EMC CX3-80
Dual redundant Controllers (Storage Processors A,B)
4 Ouput for each SP (1,2,3,4)
SUSTAINED PERFORMANCE R/W
800MByte/s (each EMC CX3-80 sys.)

GPFS implementation

- The idea of GPFS is to provide a fast and reliable (NSPF) diskpool storage with direct access (posix file protocol) from the Worker Nodes Farm using Block level I/O interface over network – GPFS Network Shared Disk (NSD) and parallel access
- GPFS is a cluster, with a SAN hardware a true full NSPF is possible (diskservers failures just decrease the theorical bandwidth but the filesystem is still avaliable)
- One single "big filesystem" for each VO could be possible (strongly preferred by users)
- GPFS is widely used at our TIER1, GPFS filesystems are directly accessible from ALL the worker node in the TIER1 FARM
- GPFS filesystem uses parallel i/o, drastically increase end optimize the disk performances compared to single filesystem (like Castor diskpool)
- In GPFS v.3.2 concept of “external storage pool” extends use of policy driven migration/recall system to/from tape storage.
- GPFS is SRM v.2 compliant using INFN STORM (Storm http://storm.forge.cnaf.infn.it/) SRM interface for parallel file systems

GPFS on SAN

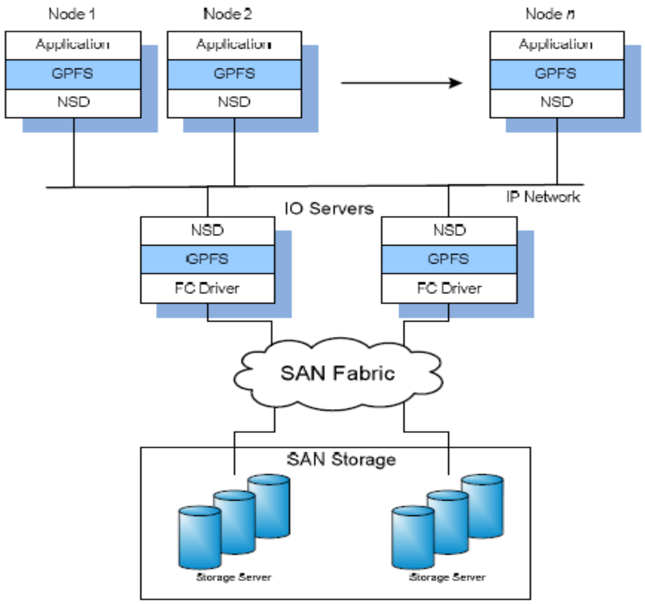
Physical layout



Farm nodes

Diskservers

Logical layout



- All diskservers accessing all disks
- All farm nodes accessing using LAN and NSD gpfs configuration
- Additional servers (i.e. front-end like gridftp) can easily be added
- Failure of a single server will only reduce available bandwidth to storage by factor N-1/N (N – number of diskservers)
- Up to 8 diskserver could be assigned to a single device i.e. the filesystem will be online as long as at 1 out of 8 servers is up
- Bandwidth to disks could be optimized using filesystem striped over different piece of hardware
- Long experience at CNAF (> 3 years), ~ 27 GPFS file systems in production at CNAF (~ 720 net TB) mounted on all farm WNs

GPFS Tape Extension

- In GPFS v.3.2 concept of “external storage pool” extends use of policy driven migration/recall system to/from tape storage.
- The "natural" choice for managing tape storage extension for GPFS is Tivoli Storage Manager (TSM also from IBM).
- External pool “rule” defines script to call to migrate/recall/etc. files to/from the external storage manager (TSM in our case).
- GPFS policy engine automatically builds candidate lists and passes them to external pool scripts.
- External storage manager (TSM) actually moves the data.
- TSM installation has been under test for more than one year at CNAF TIER1 and a LHCb production testbed is in use from Spring 2008.
  - This "Long pre-production" is due some features lacks in recall and migration policies, which is under development right now
- GPFS with an efficient TSM tape extension could be seen as a true Hierarchical Storage Manager facility.

TSM

- Agreemen with IBM to use the software until ready for full production, strong collaboration with the development team for the migration/recall optimization features
- Running Server Version 5.5, also beta version 6.1.0 client is installed for test purpose (better recall policies with "intelligent" queue and sorting optimization)
- LAN-free migration/recall to/from tape is possible. Drive should be connected to a dedicated SAN portion (Tape Area Network or TAN)
- TSM could also be easily used as a standard backup system for replacing our Legato Networker system
- TSM uses an internal database for storing filesytem metadata that could be easily duplicated. So TSM central services could be made rendundant

LHC Storage Classes at CNAF

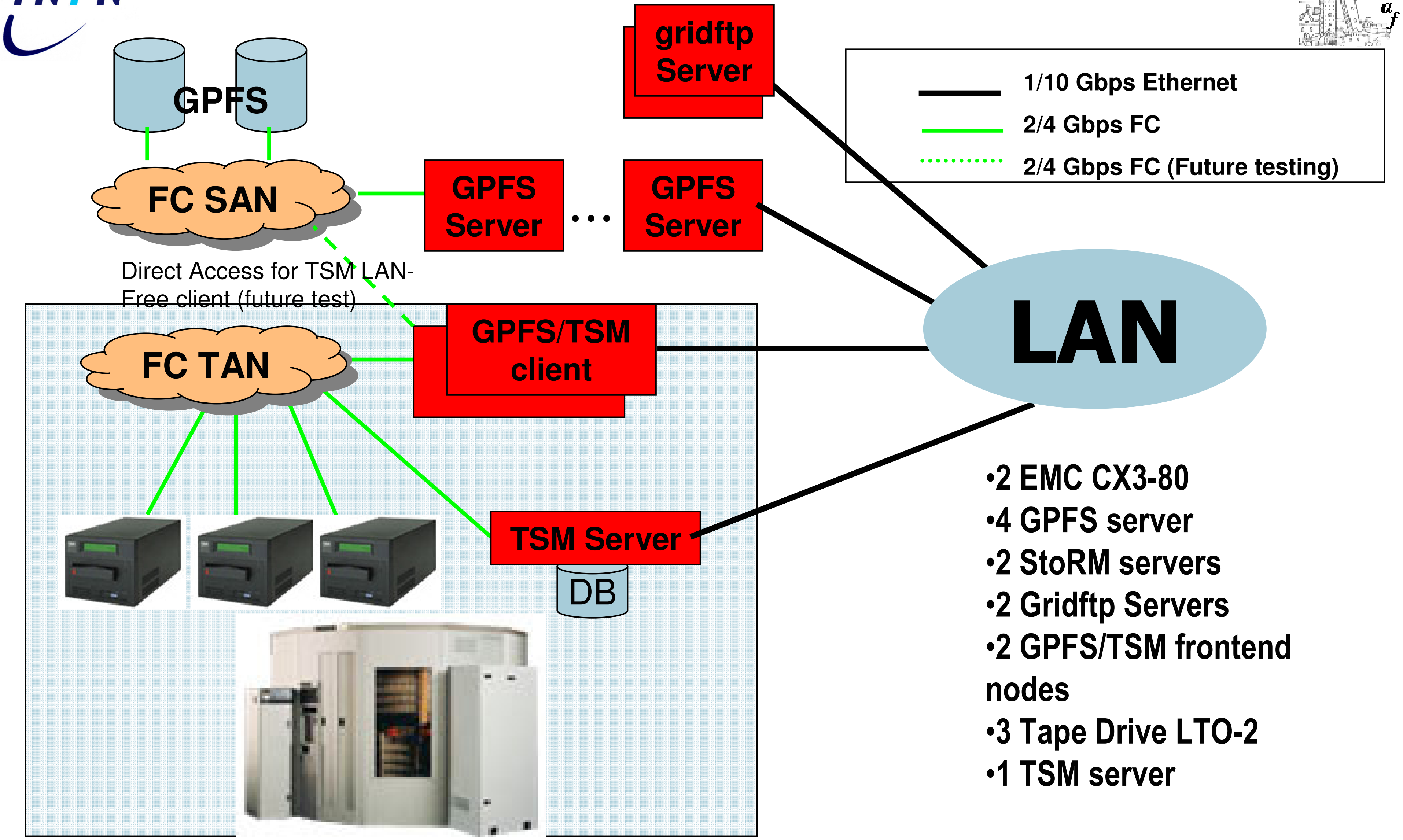
- Implementation of 3 Storage Classes needed for LHC
- Disk0Tape1 (D0T1) → CASTOR
  - Space managed by system
  - Data migrated to tapes and deleted from when staging area is full
- Disk1tape0 (D1T0) → GPFS/StoRM (in production)
  - Space managed by VO
- Disk1tape1 (D1T1) → CASTOR (production), GPFS/StoRM (production prototype for LCHb only)
  - Space managed by VO (i.e. if disk is full, copy fails)
  - Large permanent buffer of disk with tape back-end and no gc

GPFS/TSM Prototype

- 40TB GPFS File system (v.3.2.0-3) served by 4 I/O NSD servers (SAN devices are EMC CX3-80)
  - FC (4Gbit/s) interconnection between servers and disks array
- TSM v.5.5
- 2 servers (1Gb Ethernet) TSM front-ends each one acting as:
  - GPFS client (reads and writes on the file-system via LAN)
  - TSM client (reads and writes from/to tapes via FC)
- 3 LTO-2 tape drives
  - Sharing of the tape library (STK L5500) between Castor e TSM
    - i.e. working together with the same tape library
    - direct access using TAN (tape area network) for LAN free migration/recall (using TSM storage agent) will be possible (not tested yet...)

In the next slides we'll see the prototype test and the following production results

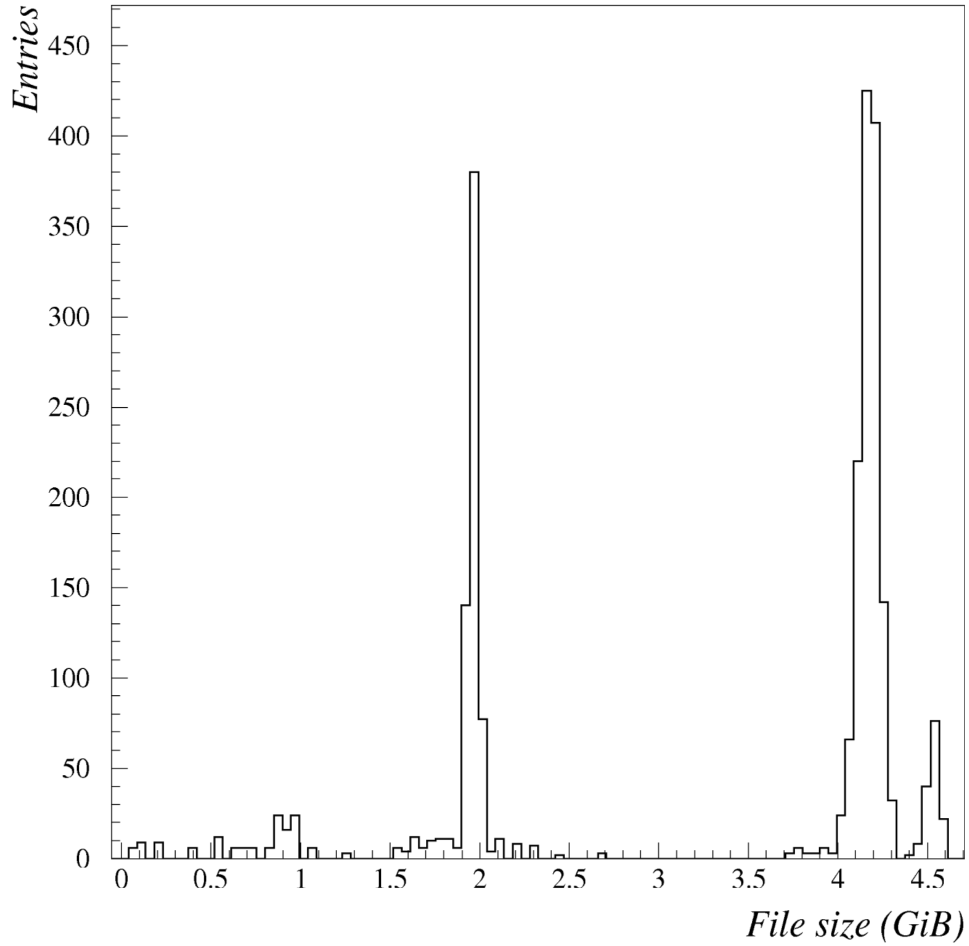
LHCb GPFS/TSM prototype and production layout



- •2 EMC CX3-80
- •4 GPFS server
- •2 StoRM servers
- •2 Gridftp Servers
- •2 GPFS/TSM frontend nodes
- •3 Tape Drive LTO-2
- •1 TSM server

GPFS/TSM Prototype LHCb Test

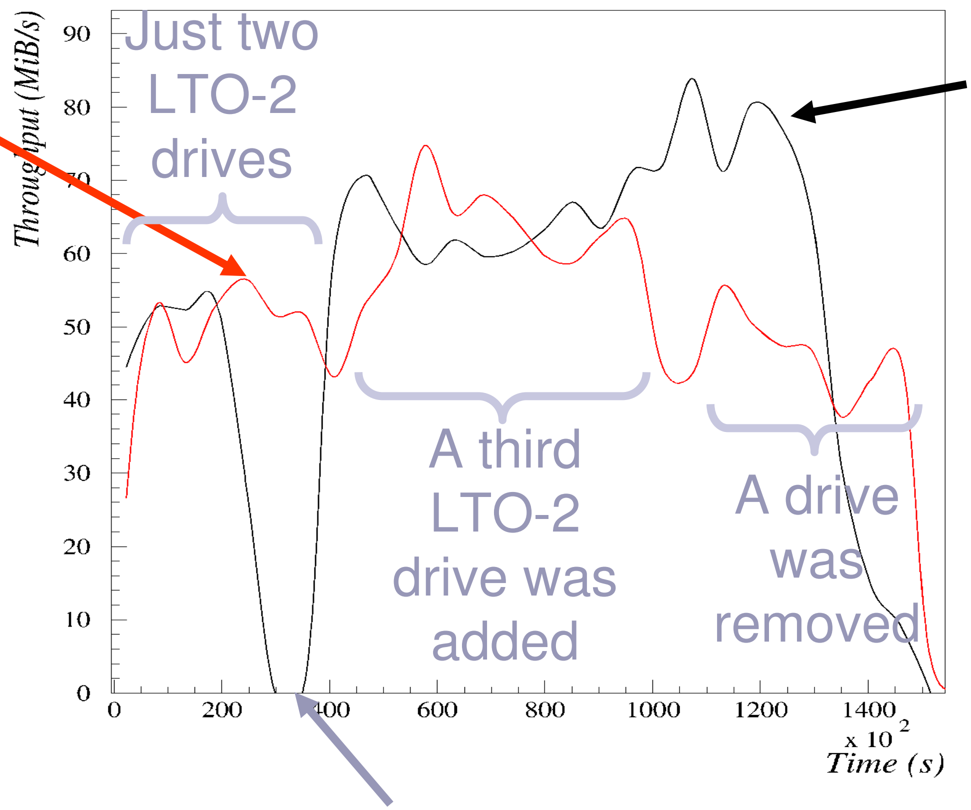
File size distribution



- Data transfer of LHCb files from CERN Castor-disk to CNAF StoRM/GPFS using the File Transfer Service (FTS)
- Automatic migration of the data files from GPFS to TSM while the data was being transferred by FTS
- *This is a realistic scenario!*
- Most of the files are of 4 and 2 GB size, with a bit of other sizes in addition
- data files are LHCb stripped DST
- 2477 files
- 8 TB in total

GPFS/TSM Prototype LHCb Test

Red curve: net data throughput from GPFS to TSM



Just two LTO-2 drives

A third LTO-2 drive was added

A drive was removed

Black curve: net data throughput from CERN to CNAF vs. time

Zero tape migration failures

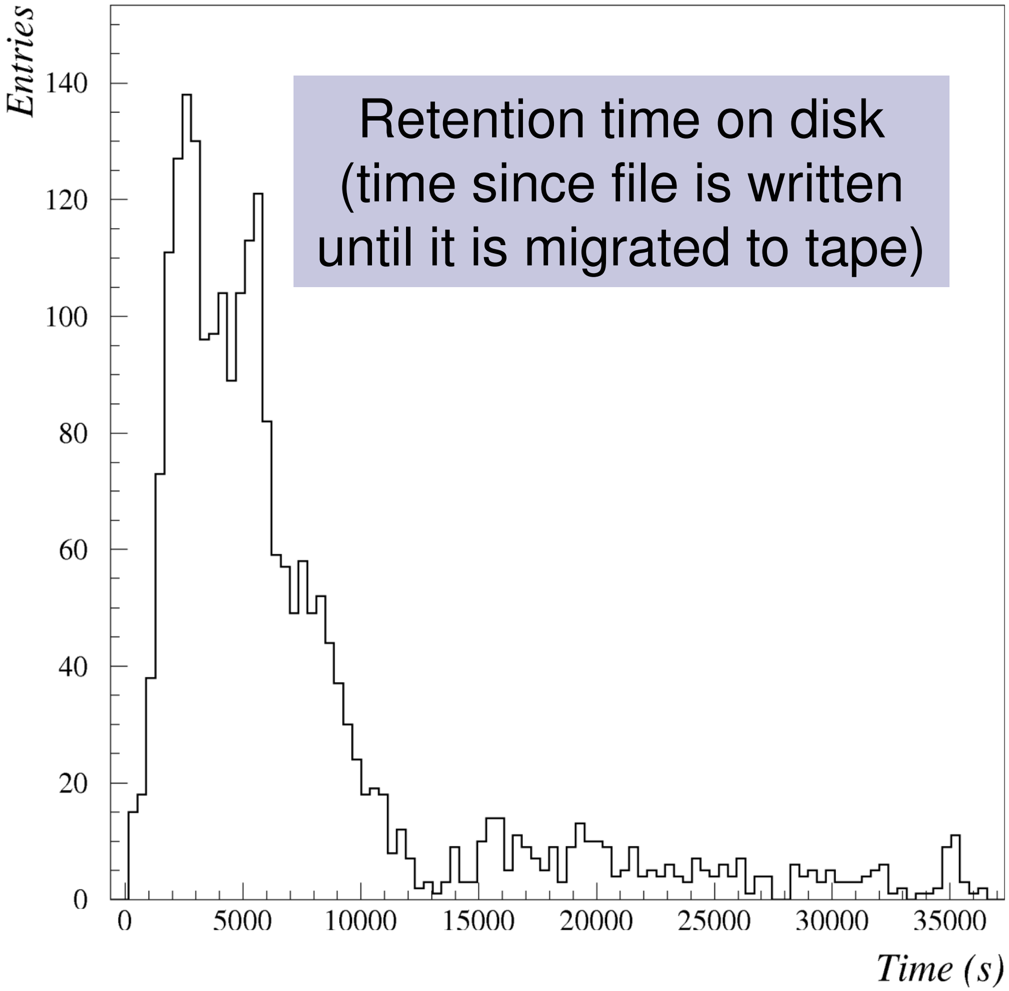
Zero retrials

8 TB in total were transferred to tape in 150k seconds (almost 2 days) from CERN

FTS transfers were temporarily interrupted

About 50 MB/s to tape with two LTO-2 drives and 65 MB/s with three LTO-2 drives

GPFS/TSM Prototype LHCb Test



Retention time on disk (time since file is written until it is migrated to tape)

- Most of the files were migrated within less than 3 hours with a tail up to 8 hours
  - The tail comes from the fact that at some point the CERN-to-CNAF throughput raised to 80 MiB/s, overcoming max performance of tape migration at that time. So, GPFS/TSM accumulated a queue of files with respect to the FTS transfers

GPFS/TSM LHCb Production

After the good results from the test phase described in the previous slides, we decide to run the prototype in production.

- 40 Tbyte of D1T1 LHCb production data successfully stored
- About 70 MByte/s sustained
- No tape migration failures detected
- A test of complete deletion of portion of the Disk Filesystem and successive full recovery from TSM tape has been made (using the TSM metadata db)
- A very promising starting!

Conclusion and "What's next"?

- This presentation contains a site report from the INFN CNAF Tier1 Storage Group activities focusing on Database, Castor, SAN and GPFS usage at our site.
- In addition the presentation briefly summarizes the promising implementation of the new GPFS/TSM prototype.
- The GPFS/TSM prototype with the SRM StoRM interface proves itself as a good and realiable D1T1 system, LHCb is still using this system in production.
- Next Steps will be:
  - A D0T1 storage class implementation of the system in collaboration with the IBM development team. Since operation of recalls becomes crucial in D0T1 systems, optimization in accessing data stored on tapes becomes of primary importance
  - Also LAN-Free migration/recall to/from the tape facilities should be carefully tested. Using the SAN/TAN for migrating and read the data between the GPFS and TSM layers could seriously improve the performance and decrease the LAN data troughput request
- Thank you for the attention!

Abstract

Title:

Mass Storage System for Disk and Tape resources at the Tier1.

Abstract: The activities in the last 5 years for the storage access at the INFN CNAF Tier1 can be enlisted under two different solutions efficiently used in production: the CASTOR software, developed by CERN, for Hierarchical Storage Manager (HSM), and the General Parallel File System (GPFS), by IBM, for the disk resource management. In addition, since last year, a promising alternative solution for the HSM, using Tivoli Storage Manager (TSM) and GPFS, has been under intensive test. This paper reports the description of the current hardware and software installation with an outlook on the last GPFS and TSM tests results.