



# CASTOR CNAF TIER1

Barcelona PIC 18-19 Nov. 2004

Ricci Pier Paolo [pierpaolo.ricci@cnafe.infn.it](mailto:pierpaolo.ricci@cnafe.infn.it)

Lore Giuseppe [giuseppe.lore@cnafe.infn.it](mailto:giuseppe.lore@cnafe.infn.it)

[castor-support@cnafe.infn.it](mailto:castor-support@cnafe.infn.it)

# TIER1 CNAF PRESENTATION



- Hardware and software status of our CASTOR installation
- Usage from LHC experiment of our installation
- Problems arisen
- Comments (next generation stager: Castor-2)

- At present there are 3 people at CNAF involved in administering our CASTOR inst. and front-ends:
- Ricci Pier Paolo** (50% also activity in SAN/NAS HA disk storage management and test, Oracle adm., KVM systems, Electric Power and remote PDU of our Computing Room and in purchase HW)
- Lore Giuseppe** (50% also activity in ALICE exp. as Tier1 reference, SAN HA disk storage management and test, managing Grid frontend to our resources)
- Vilucchi Elisabetta** (30% new employee also involved in Oracle and RLS development and adm. and SAN disk storage management and test)

# MENPOWER (2)



## Please notice:

- ✦ We don't have people full time involved in only our Castor Management (no R&D)
- ✦ At present we also don't have a TIER1 CNAF person working with the Development Team at CERN, needed for 1st level assistance and specific development of the code for our installation (i.e. LTO-2 certification with IBM drivers, specific command certification like the rfstage, stager patches...)
- ✦ We will focus on the previous point: we need soon at least 1 person working at Cern under coordination of the Development Team to fulfill with priority TIER1 CNAF requests

# HARDWARE STATUS



Our old STK L180 library and LTO-1 drives has been dismissed (repacked on LTO-2) from our castor installation so at present our CASTOR system is:



1 STK L5500 SILOS  
partitioned with 2 form-factor  
slots

About 2000 slots LTO-2 form  
About 3500 slots 9940B form



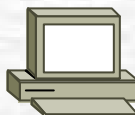
6 LTO-2 DRIVES with  
2Gb/s FC interface

1300 LTO-2 Imation  
TAPES



2 9940B DRIVES with  
2Gb/s FC interface

500-800 9940B TAPES



Sun Blade v100 with 2  
internal ide disks with  
software raid-0 running  
ACSL 7.0

# HARDWARE STATUS (2)



8 Tapeservers, 1U Supermicro 3 GHz 2GB with 1 Qlogic 2300 F.C. HBA, STK CSC Development Toolkit provided by CERN (with licence agreement with STK) ssi, tpd daemon and rtcpd.

The 8 tapeservers are direct connected directly with the FC drive output:

DRIVE LTO-2 0,0,10,0	->	tapesrv-0.cnaf.infn.it
DRIVE LTO-2 0,0,10,1	->	tapesrv-1.cnaf.infn.it
DRIVE LTO-2 0,0,10,2	->	tapesrv-2.cnaf.infn.it
DRIVE LTO-2 0,0,10,3	->	tapesrv-3.cnaf.infn.it
DRIVE LTO-2 0,0,10,4	->	tapesrv-4.cnaf.infn.it
DRIVE LTO-2 0,0,10,5	->	tapesrv-5.cnaf.infn.it
DRIVE 9940B 0,0,10,6	->	tapesrv-6.cnaf.infn.it
DRIVE 9940B 0,0,10,7	->	tapesrv-7.cnaf.infn.it



# HARDWARE STATUS (3)



castor.cnaf.infn.it Central Machine 1 COMPAQ DL360  
raid1 R.H. 7.2 Machine running all CASTOR 1.6.1.3 service +  
ORACLE 9.1 database (Nsdaemon, vmgrdaemon, Cupvdaemon,  
vdqmdaemon, msgdaemon, rfiod, ctpdaemon and test stager), will  
be soon migrated in 2 different machines/HW with the same name

castor-1.cnaf.infn.it Monitoring Machine 1 DELL 1650  
R.H 7.2 Machine running monitoring CASTOR service (Cmon  
daemon) , ORACLE Recovery Manager Database (needed for  
LEGATO backup of the db) and NAGIOS central service for  
monitoring and notification. Also contains the command rtstat e  
tpstat that are usually runned with the -S option over the  
tapeserver

**Stagers with diskserver:** 1U Supermicro 3 GHz 2GB with 1 Qlogic 2300 F.C. HBA accessing our SAN and running Cdbdaemon, stgdaemon and rfiod

disksrv-1.cnaf.infn.it ATLAS stager with 2TB locally

disksrv-2.cnaf.infn.it CMS stager with 3.2TB locally

disksrv-3.cnaf.infn.it LHCb stager with 3.2TB locally

disksrv-4.cnaf.infn.it ALICE stager with 3.2TB locally

disksrv-5.cnaf.infn.it FREE stager (archive purpose LVD,alice TOF etc..)

**Diskservers:** 1U Supermicro 3 GHz 2GB with 1 Qlogic 2300 F.C. HBA accessing our SAN and running rfiod

At present we are testing in production HA configuration where the rfiod daemon runs over a virtual IP that can be shared over a failover domain of 2-4 machines accessing the same SAN area.

LHCb has 16TB of staging area over one Virtual IP

ATLAS has 4TB of staging area over one Virtual IP



# HARDWARE STATUS (5)



3 Storage Element front-end of CASTOR

castorftp.cnaf.infn.it (default)

castorftp-1.cnaf.infn.it (LHCb)

castorftp-cms.cnaf.infn.it (CMS)

castorgrid.cr.cnaf.infn.it (load balaced)

Installation of SRM v1 (on going)

- Substituted 1 rpm (E. Knezu)
- Mapping certificate – user grid
- Mapping user - (stager,diskpool)

# CASTOR USAGE



At present we have 4 LHC experiment that are using castor, ALICE, ATLAS, CMS, and LHCb.

Also 2 other HEP experiment will start soon accessing our tape capability using castor as a backup for their data on the disk media (VIRGO 14TB, BABAR 30TB, CDF 30TB)

The access to the castor system is

- 1) Grid using our SE frontends (from WAN)
- 2) Rfio using castor rpm and rfio commands installed on our WN and UI (from LAN)

# CASTOR USAGE (2)



LTO-2 Space used after about 9 months of activity (1200 TAPES total)

ALICE	26 TAPES	(2 marked read-only)
ATLAS	27 TAPES	(10 marked read-only)
CMS	54 TAPES	(all read-only ½ space)
LHCB	68 TAPES	(2 marked read-only)

Only the 15% of the total space was effectively used by the experiments:

- 1) As TIER1 storage we offer “pure” disk as primary storage over SAN (preferred by the experiments)  
(GSIftp,nfs,xrootd,bbftp,GPFS ....)
- 2) The lack of an optimization in parallel stage-in operation (pre-stage) and reliability/performance problem arisen in LTO-2 give in general very bad performance when reading from castor so experiment like cms wants pure disk(50TB) at present

# CASTOR USAGE (3)



In general we archive

- 1) Good performance when writing in the staging area and from stage to tapes (2 parallel streams on tape give about 60MB/s)
- 2) General good reliability on the stager service (since experiment knows the single stager problem of more that 300.000 entries and try to minimize the number of files written)
- 3) Bad reliability on LTO-2 drive when writing and reading. This results in tapes marked readonly or disabled when writing and in locking or failure when trying to stage-in files in random order (using rfstage of ordered files and streams operation gives great improvement but need big staging area)

# CASTOR USAGE (4)



ALICE: Decided to a staging-area/tape-area ratio of 1. So the stage-in needed in the next analysis/processing phase will be negligible

ATLAS: Didn't coordinate a good and clear production/analysis plan (only 2TB of staging area for 25x200GB tapes) and starts analysis on files that where all on tape. Our LTO-2 system was struck

LHCB: Plan a initial staging-area/tape-area low ratio when in production (2 months) but now they need to analyze the whole data. Coordinating with us we temporary increment the ratio to 1 and trigger an optimized stage-in of all the tapes. With 3 parallel tape to disk streams (dedicated drives) we read all with a total bandwidth of 60MB/s without "major" problems (apart the duplicated segment bug). Also they decided to use "big" files to overcome writing problems

CMS: They are trying to compact their files in "big" files for the next data challenge but they decide to use pure disk storage element instead of increasing the staging area of castor



# PROBLEMS



We notice a direct connection on the use of our LTO-2 drives and the increase of the error rate:

- 1) The number of files on a single tape
- 2) The size of the file to be written/read (KB)
- 3) The number of mount/dismount rate on single tape/drive

In general seems that a big number of rewind/fskip operations give the higher error rate in chksum (writing) or no sense (reading) (locate,fskip)

Also a great number of mount/dismount result in problems when loading the tape (tape positioned but not loaded in the drive) or tape blocked in drive.

STK/IBM cannot give us solution at this level other than changing the drives (all 6 were replaced during the past 9 months!) and suggest to use their drivers (not castor compatible...)



## PROBLEMS (2)



Also there is a common situation where the locate (or skipfile) function doesn't return (usually in a read operation) therefore the tpd daemon remains "hanged" forever.

Usually functions childs of tpd daemon posovl or rlstape are unkillable (probably for FC timeout)

The drive remains in an "open" condition, doesn't accept scsi commands (i/o error, device open or simply hang) and cannot be unloaded by acsls. Usually some reboots of the tapeserver can fix the condition but sometimes power-off of the drive is needed

## PROBLEMS (3)



The choice of migrate to 9940B was taken as we think that the different error types and rate on LTO-2 cannot be solved at this level.

But we cannot trade-in our LTO-2 drives (6) and tapes (1200 out of 2000 slots) and we must use them in some way.

LTO-2 drives work and perform well when heavily doing sequential operation on big files (100MB or more) with a small amount of mount/dismount every day

This is why we strongly suggested experiments to use big file (zipped) and to coordinate with us the analysis of the data for triggering an optimized stage in (and therefore increasing the stage buffer)

## PROBLEMS (4)



We implement the rfstage command for internal use only and some script to obtain info about segment (tape and position)

Giving to the rfstage command a LIST OF ORDERED FILE in a SPECIFIC TAPE the users will soon request for analysis we were able to drastically reduce the error rate when reading and increase the performance of a factor of 2-10 or more

Also we need in the previous write operation to be sure that data that has been written sequentially (usually in the same directory) are phisically stored in the same tapes.

This can be done limiting the streams when writing and understanding the stager migration policy (the migration seem to be done using a size order not a FIFO one and cannot be scheduled at this level)

# COMMENTS



We think that the next generation stager (castor2) and vdqm should be able to prevent most of the errors arisen in our TIER1 installation.

Some points should be considered regardless of the site specific hardware:

- 1) Drives performs better in streams operations ("big" files and heavy migrations operations)
- 2) A big number of mount/dismount usually can be overcome by policies
- 3) Rewind and locate/fskip operations also should be reduced as much as possible
- 4) Data written in the same period (productions) usually will be re-read (analyzed) all in the next lcg phase

## COMMENTS (2)



tpread, tpwrite: The currents tpread and tpwrite has some limits? (Number of files written or read in a single operation, size of a single stage stream...)

It's worth thinking of a "dinamic" tpread tpwrite operation where segment are added "on the fly" of another operation currently accessing the drive (see below)

Vdqm: Normally when another request in queued for a tape that is currently accessed, it it not unmounted/remounted (see PIC) and a corrisponding new tpread/write operation is triggered. Is it possible (and worth working) add a optimization in the vdqm queue list for the ORDER of the segment (this will save locating time in particular for rewind if posovl is forced to use skiptpff)?



## COMMENTS (3)



stgdaemon: There is a limit in the number of files a user can stage-in/stage-out in a single operation or in multiple operation? It will be useful to limit this number or implement a some sort of time-out period during which the stager "collect" requests before passing them all together (after optimization) to the vdqm (so limiting the reqID to few optimized operations)?

Also how is managed the STAGE-IN status after down of the stgdaemon (crash or kill)?

How is possible to force the migrator to write specific data to specific tape lists in specific order?

It could be useful to link the new stager (since it will be centralized) to the nsdaemon so the nsis -l (or other grid tool) could include an S "flag" if the file is already present in the staging area? (like the M flag)



## COMMENTS (4)



stgdaemon: reading linear tape in forward streams of big amount of data helps in performance and reliability of tapes (limiting the number of start-stop and physical locate operations). An optimizer at stager level which orders and compact requests will improve both. (In our LTO-2 hardware this will means difference between an usable or unusable system)

Also the problem of "small" file is emphasized in LTO-2. We had great problem during CMS datachallenge (small file: 20000 segment in 200GB with most size of 10-100KB)

This kind of data give us lots of problem writing and are almost unreadable without rfstage optimized prestaging (single request, stream reading)

## COMMENTS (5)



Rfstage (admin only): Could be improved? (accepting whole CASTOR directory and optimizing order).  
The argument list could be expanded? (limit to about 1000 HSM file name)

Stagein: Will be dismissed in Castor2?

Repack: Could be useful to modify this command to leave administrator the options to stage-in files with the castor filenames so they can remain on the staging area? (not only with the temporary names that are deleted from the staging area)

## COMMENTS (6)



Rfiod: How currently the stager can manage or recover from rfiod (on diskservers) failures (timeout,etc...) and could it manage failover with IP aliases during single file reading/writing? And the next generation stager?

The new stager service could be HA? The vdqm? The nameserver? The vmgrdaemon? (they are all the central castor service)