Understanding and improving the tape performance at PIC Tier-1 after STEP'09: the CMS case

Josep Flix *
PIC, CIEMAT, Barcelona, Spain

EGEE'09 Conference in Barcelona
21-25 September 2009

* *Thanks to PIC Storage Team: i.e. E. Acción, G. Bernabeu, M. Caubet, F. Martínez*

Outline

- The MSS Setup at PIC Tier-1 center
- Tape performance results during STEP'09 at PIC, for CMS:
  - Configuration and results
  - Lessons learned during STEP'09
    - Increasing # of recall pools
    - Low read/write drive rates (LT04)
    - Tuning dCache Pool Costs
    - CMS pre-stage of data: LAN/WAN accesses
- July'09 LT04 tapes read tests at PIC
- How to improve reads/writes from/to MSS system:
  - From VOs p.o.v → The CMS case
  - From Site p.o.v → MSS configuration

MSS Setup at PIC

- At PIC, two co-operating systems: **dCache** and **Enstore**
  - **dCache** visualizes storage clusters in a single name space
  - **Enstore** provides distributed access and management of data stored on tape
  - Single dCache instance for the 3 LHC VOs supported @ PIC (ATLAS, CMS, LHCb)

- At PIC, two tape robots available:

STK SL 8500

IBM TS3500

Tape Drives

STEP'09

7 TD 9940B (STK)
7 TD IBM LTO3 (IBM)
4 TD IBM LTO4 (IBM)
5 TD STK LTO4 (STK)

AS TODAY

7 TD 9940B (STK)
7 TD IBM LTO3 (IBM)
4 TD IBM LTO4 (IBM)
12 TD STK LTO4 (STK)

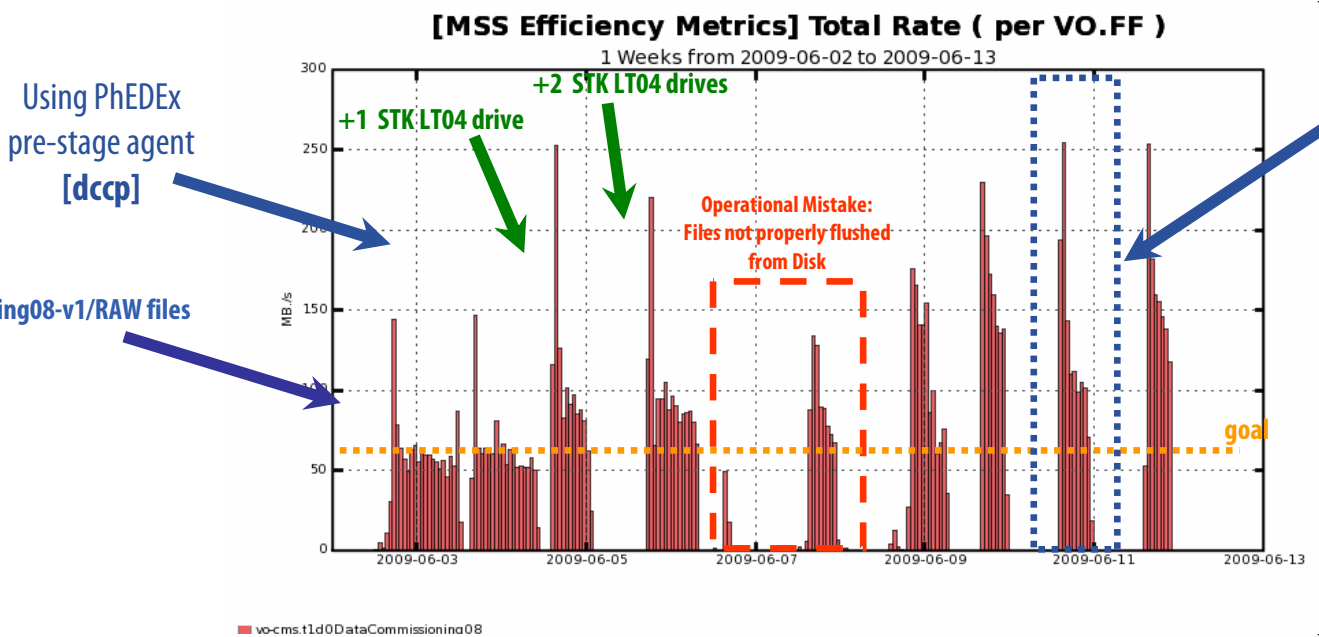
CMS STEP09 reprocessing tests @ PIC I

- Simultaneous tests of **pre-staging** and **rolling processing** in STEP09 (2-week period):
  - Pre-stage + re-process programmed from June the 2nd to 14th
  - Special processing of data **not-prestaged** in 15th June
- At PIC, chosen dataset (~50 TBs, only **RAW files**) located on two **different** tape robots:
  - 90% (10%) of files in STK (IBM) tape robot → All RAW data stored in **LT04** tapes
- **~4.2 TBs/day** data pre-staged (~1500 files/day), and 120 (20) Tapes/day used in STK (IBM)
  - **Dataset RAW + RECO files stored on same Tapes** (*t1d0_data* FF) - **mixture**
  - The average number of files read per tape/day ~10 **[1/3 of a tape, approx.]**
- Number of STK drives increased during the test:
  - 2(beg.)→5(end) drives in STK robot & 4 drives in IBM robot

CMS STEP09 reprocessing tests @ PIC II

- All CMS pools enabled as '**recall pools**' (read buffers from MSS system)
  - STEP'09 → 13 disk pools setup as recall pools
  - This avoid possible network bottlenecks
- CMS does **not have** a Workload Management integrated prestaging system atm:
  - Used 'PhEDEx pre-staging agent', with site+MSS specific pre-staging implementations (dccp/gfal)
  - **Special configuration @ PIC**: the agent was running for both **LAN** and **WAN** tape recalls
- Pre-stage tests on top of prod. activities: PIC had **CRUZET** and **MWGR** custodial responsibilies and had other STEP09 tests. Remarkably stable site with sizable multi-VO activity (w/ATLAS, mainly)

STEP09 pre-staging tests @ PIC: recall rates I



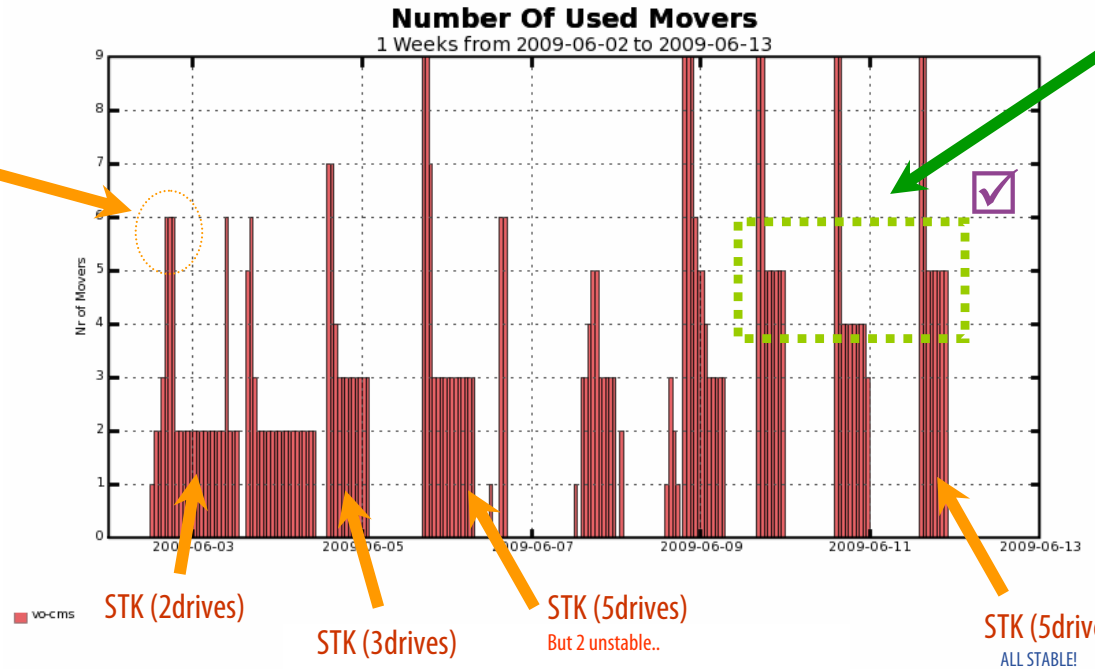
Using PhEDEx pre-stage agent **[dccp]**

/Cosmics/Commissioning08-v1/RAW files

1 day test using PhEDEx pre-stage agent **[gfal]**

RECALL RATE

Total Volume = 41.3 TBs



IBM (4drives)
STK (2drives)

Installed STK LT04 drives stable at the end

STK (2drives)

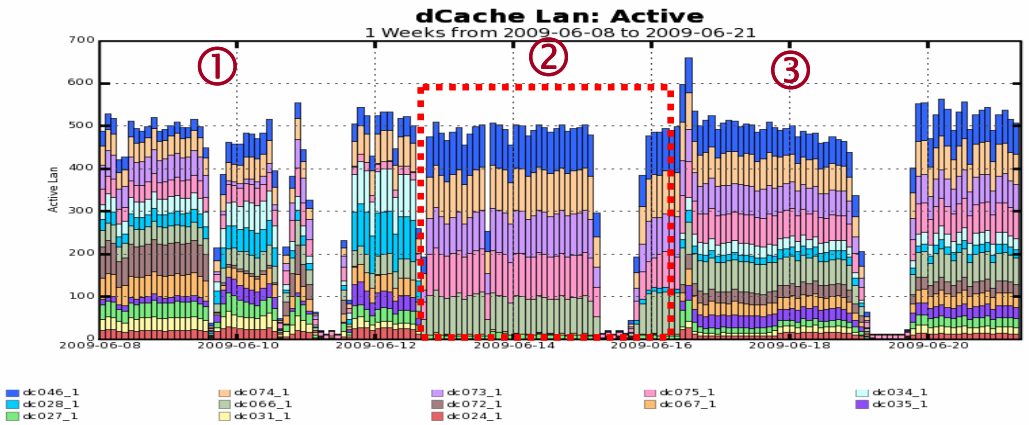
STK (3drives)

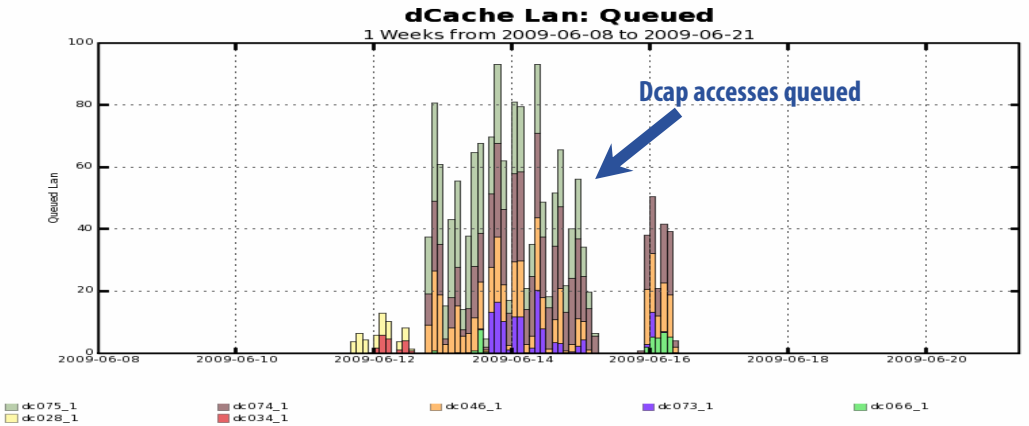
STK (5drives)
But 2 unstable..

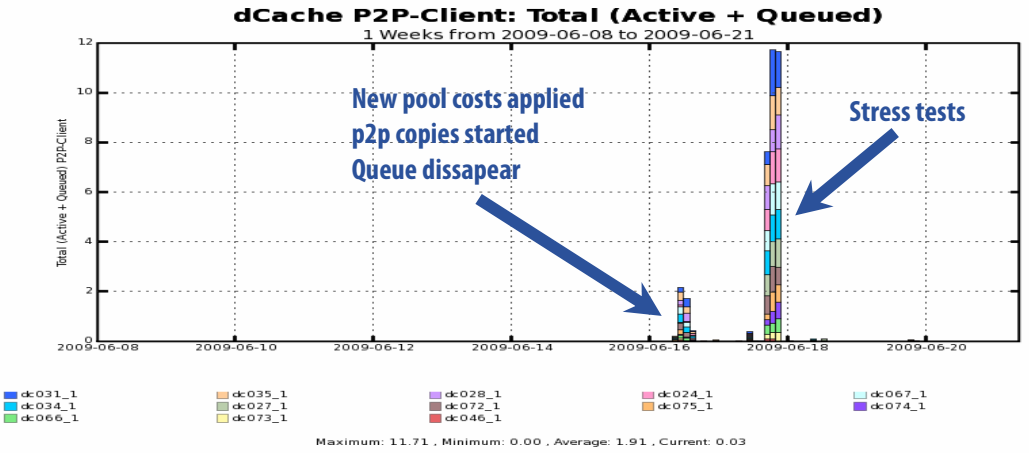
STK (5drives) + IBM (4drives)
ALL STABLE!

USED DRIVES

STEP09 pre-staging tests @ PIC: recall rates II



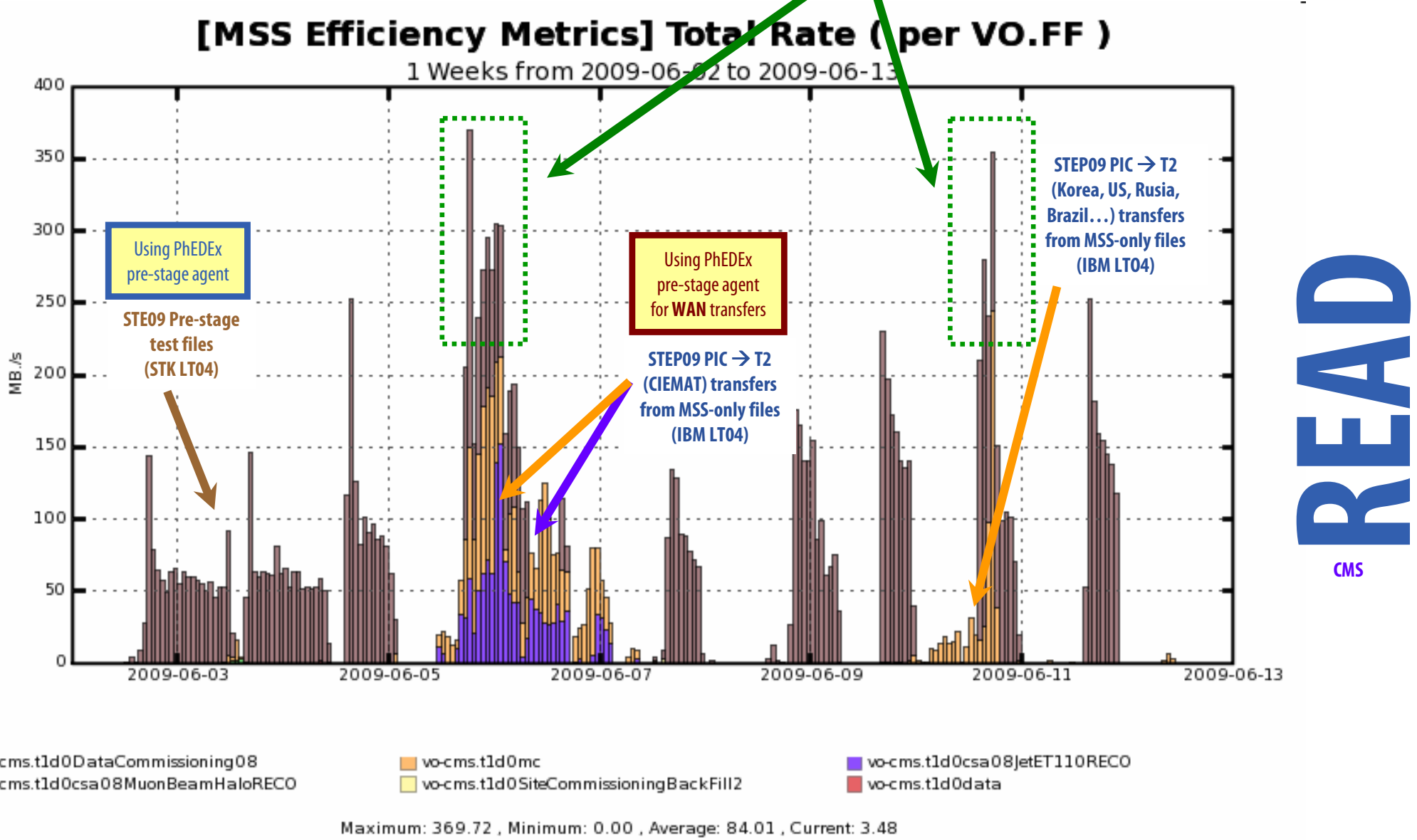




- In region ① we were using **all CMS pools** (13) for tape recall
- Pools were filling everyday in an homogeneous way
- A new bunch of recalls entered late 11/06, and during those file recalls the pools usage changed (based on pool costs)
- Therefore, only **5 pools** were used, and then all the dcap requests were queued (region ②)
- Those queued requests did not fired pool-to-pool copies and then **all the queued requests were dying by timeouts** (affecting SAM tests, Job Robot jobs, Production jobs...)
- The **only remarkable incident during STEP'09**
- We modified the pool costs, so p2p copies were fired for those requests queued. The effect is clearly seen in region ③
- The problem was only affecting CMS and was seen by the CMS SAM tests. So, it was useful information about a site problem
- We did a **stress test** well after new config was applied and **validated it**.

STEP09 pre-staging tests @ PIC: recall rates III

[CMS]: sustained observed READ rates at 250-300 MB/s

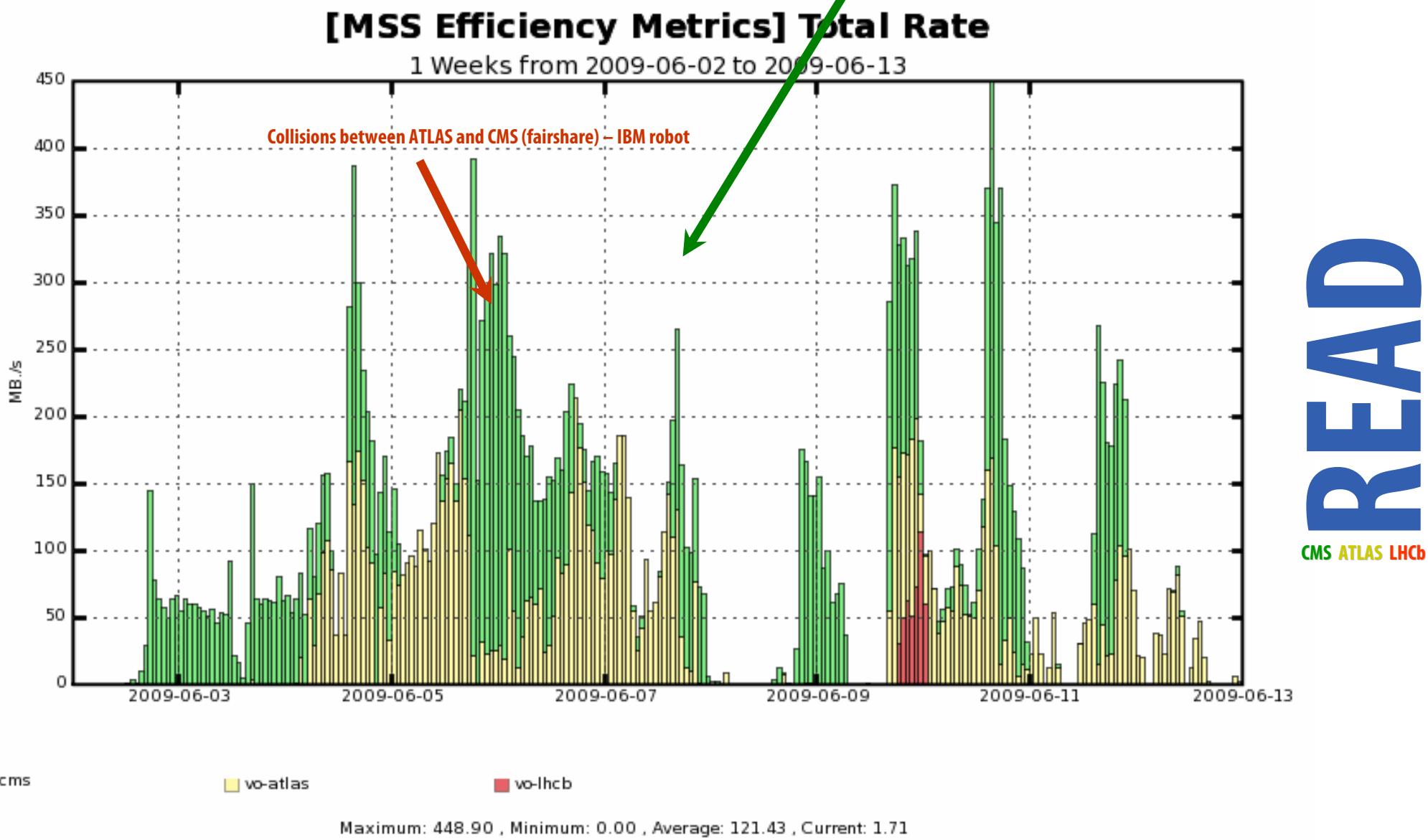


READ

CMS

STEP09 pre-staging tests @ PIC: recall rates IV

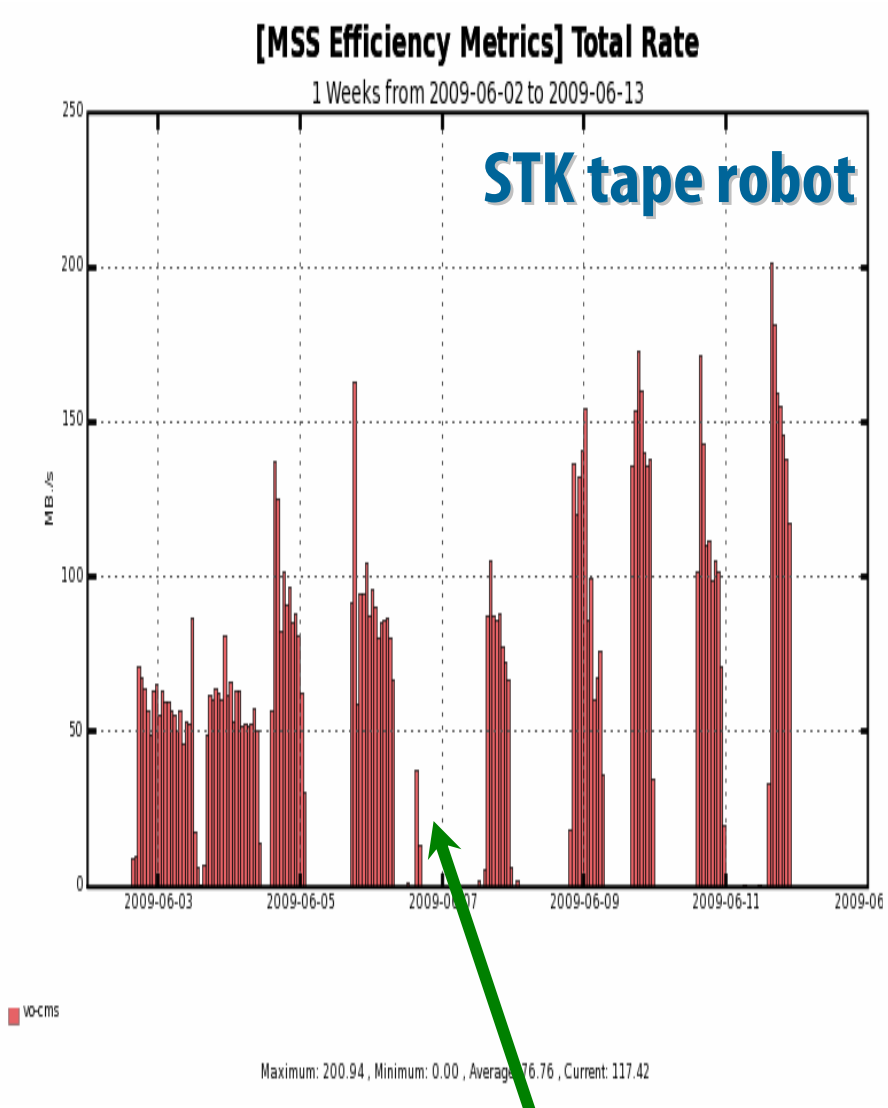
High CMS load on Tape system,
as compared to other LHC-VOs

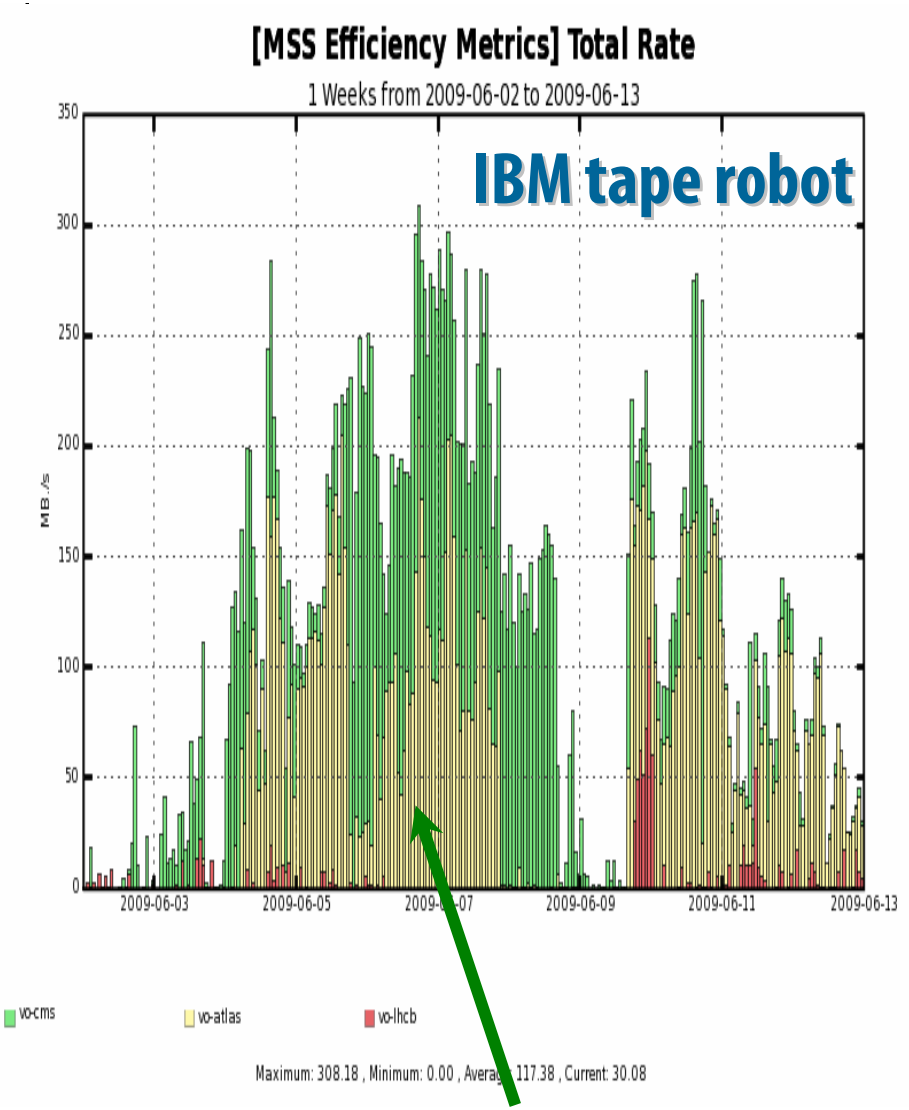


READ

CMS ATLAS LHCb

STEP09 r+w rates from/to MSS (all LHC VOs)





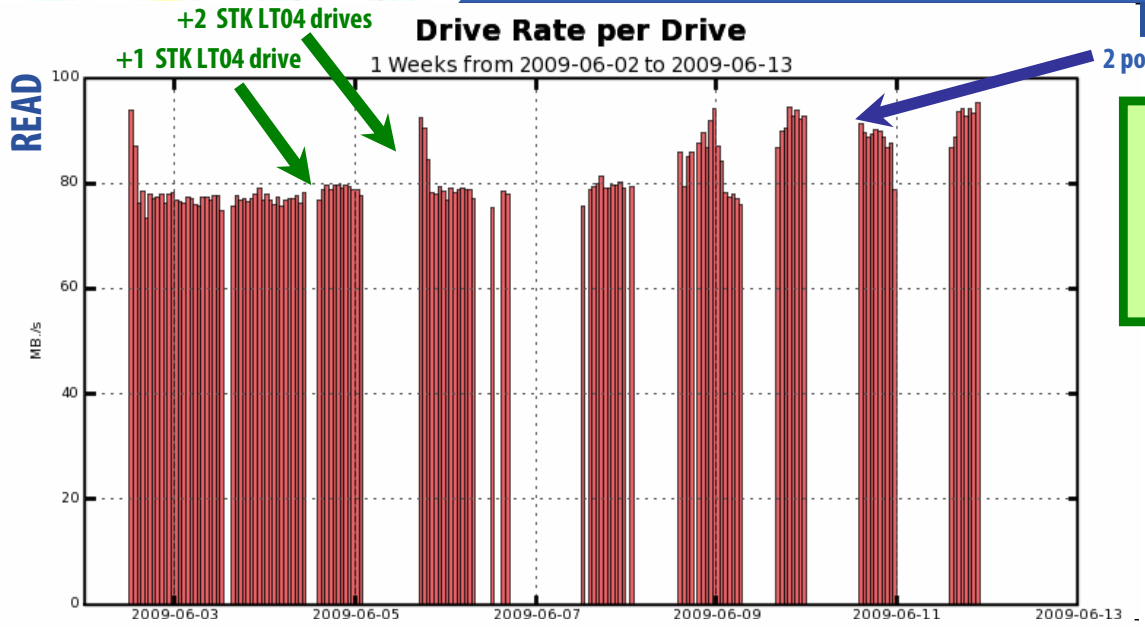
READ+WRITE

CMS ATLAS LHCb

CMS STEP09 pre-stage tests data located in the STK robot (LT04)

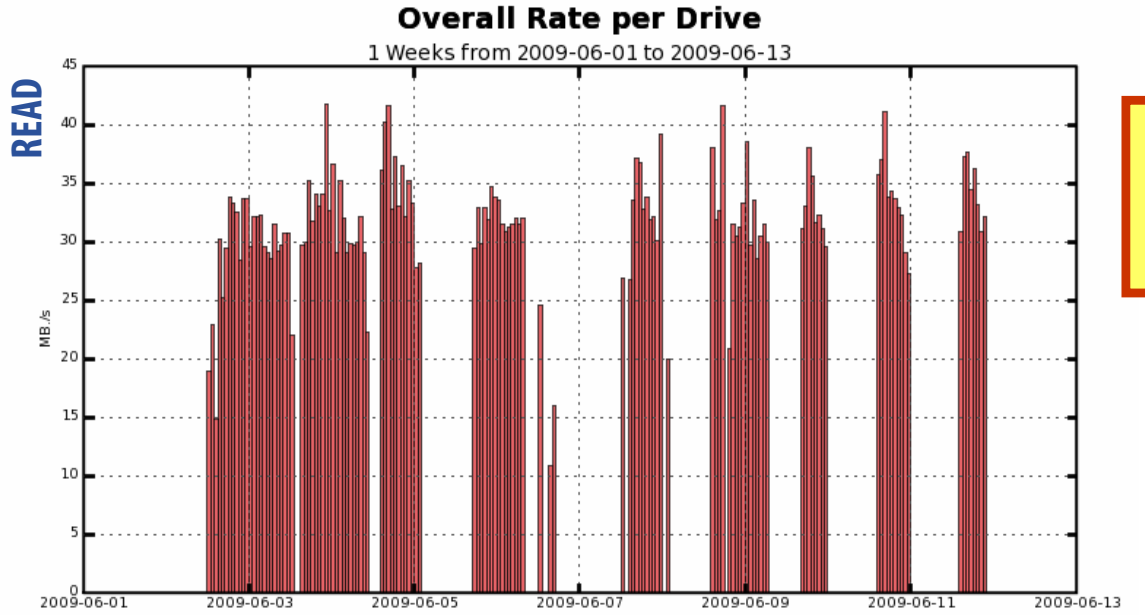
Other CMS STEP09 reads/writes went to the IBM robot (LT04). ATLAS used LT03 and LT04 drives

STEP09 pre-staging tests @ PIC: drive rates



2 powerful Tape Servers → 2 drives @ 110 MB/s/file; others @ 80 MB/s

File read rate on drives:
3 drives @ ~80 MB/s/LT04drive
2 drives @ ~120 MB/s/LT04drive



- Large dead times in **LT04** due to:
  - Tape finding
  - Tape Mounting
  - **LT04 file seeks (as big as 1')!!!!**
- **60% rate reduction using LT04…**

Overall rate from drives:
30-35 MB/s/LT04drive
(file seeks dominate the overall rates/drive)

Big file seeks on LT04

- Same test done with **T9940B** tech.:
  - Drive rate ~28 MB/s
  - Overall rate ~25 MB/s
- **Small rate reduction (10%)**

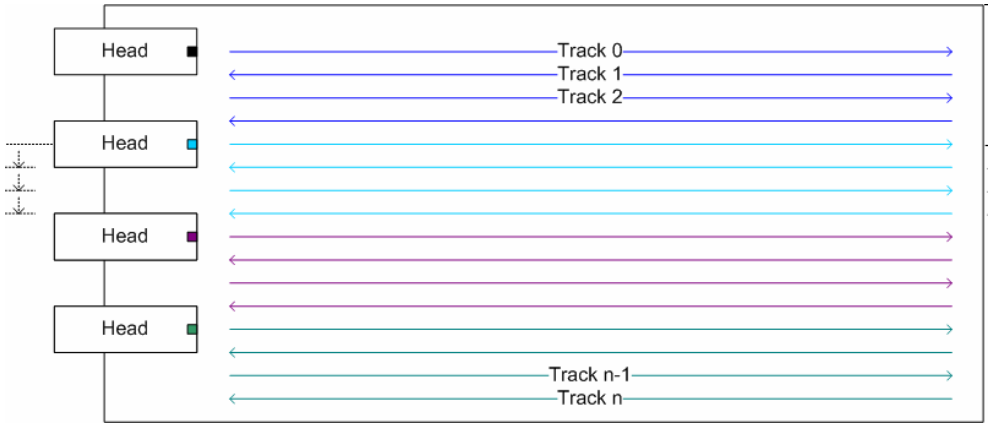
*** Same effect seen at other T1s with LT04 tapes as well…**

Lessons learned during STEP'09

- **Pro:** All LHC VOs supported at PIC passed the STEP'09 metrics
- Extensive usage of PIC MSS systems, but ATLAS/CMS files mainly placed on different robots:
  - →**Con:** Low VO-interference. Sharing of drives between experiments to be well tested
- **Con:** Low tape drive read throughputs 30-35MB/s compared to LT04 benchmark (~120MB/s)
- **Con:** Low tape drive write throughputs: LT04 drive rates ~60 MB/s; Overall rates ~25 MB/s
  - no *write pool buffers* enabled on disk pools during STEP'09 (see later)
- **Pro:** having sufficient disk pools for MSS files recall
  - But pools are then exposed to WAN transfers, accesses from WNs + Tape recalls…
    Pool costs need to be well adjusted to allow *p2p* copies when pool queues are full
- **Pro:** Importance of enabling pre-staging mechanism for WAN transfers
  - Tier-2s can be interested to get old data from Tier-1s; chaotic access affects to MSS performance.
    CMS PhEDEx pre-stage agent for WAN transfers controls the organized tape access.

LT04 Tapes: file seek times

- One tends thinking that a tape is written sequentially…
- However, to increase write performance, tape drives use more than 1 head, and to increase capacity there are more tracks than heads (and tracks are written in *serpentine* mode)



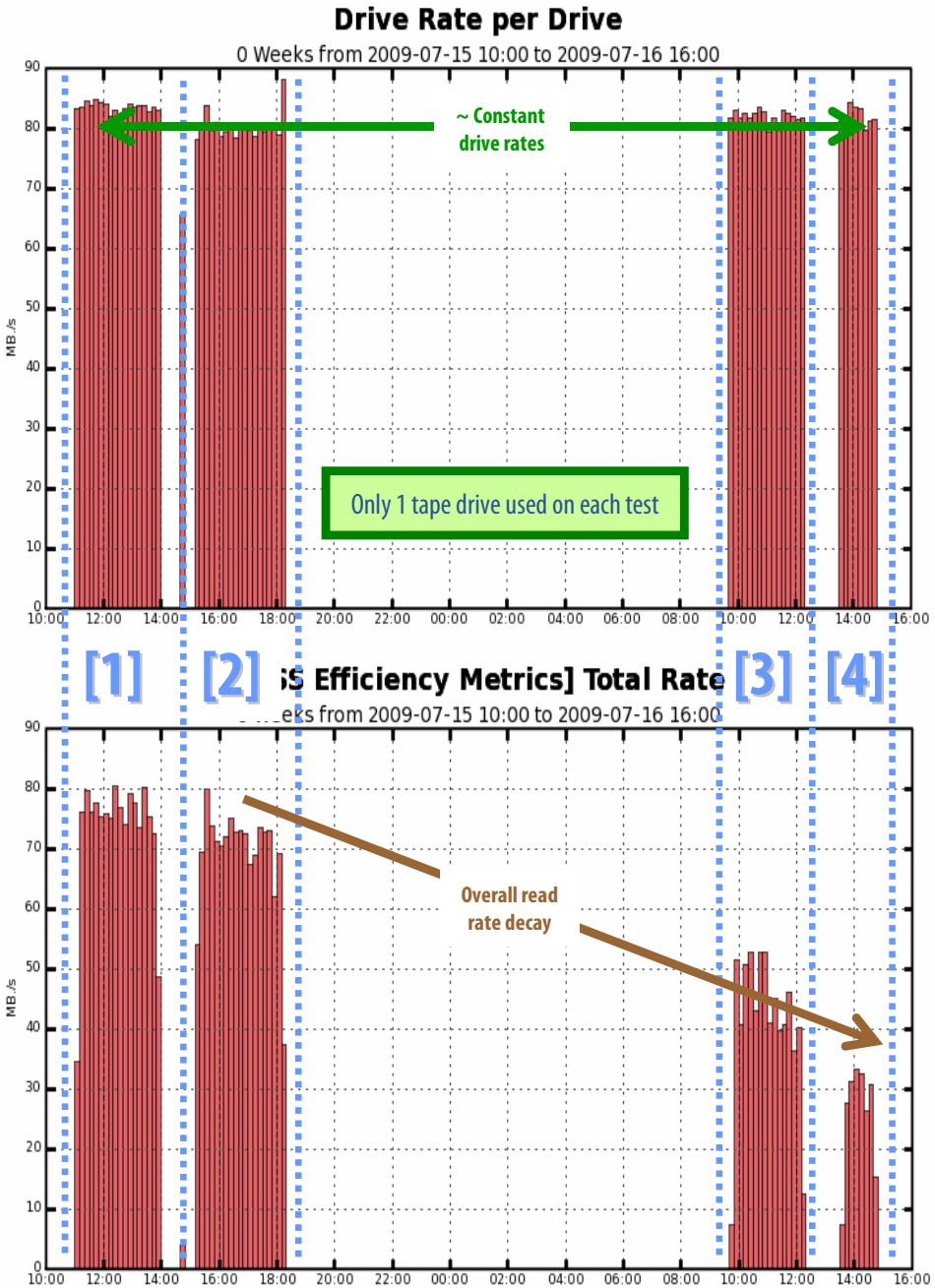
General specifications				
Tape type	Native capacity	Total tracks	Tracks written per pass	Passes to write entire tape
LTO-1	100 GB	384	8	48
LTO-2	200 GB	512	8	64
LTO-3	400 GB	704	16	44
LTO-4	800 GB	896	16	56

- Then, files are not positioned in tape following the *"fseq"* order (writing files order)
  - *fseq10* file can be near tape start than *fseq3* file...
  - Consecutive file reads → **no file seeks!**
  - But, files sufficiently separated on tapes can be **random placed**
  - The manufacturer's average seek time from beginning of tape to a random file is **1m00s**
  - From one random spot to another random spot should then be 2/3 of this, or **~40s**

LT04 performance tests [July '09] I

- To estimate the *file seek* impact, we made some tests on LT04 tapes at PIC:
  - As there are 56 "wraps" or changes in direction in an entire 800GB LT04 volume, skipping more than 10-14 GB worth of files, one is essentially moving to a random point
- We planned/did 3 tests with a CMS LT04 tape:
  - Read the complete tape → **best case**
  - Alternate file read, based on "location cookie" (fseq 1,3,5,7) → **intermediate case**
  - Alternate file read, with file seeks > 12 GBs in distance → **worst case**
- The CMS LT04 tape selected had 278 files, <file size> ~ 2.75 GBs
- Compare **drive rates** (file read rate on the drive) with **overall rates** (that takes into account file seeks, tape mounts,...)

LT04 performance tests [July '09] II



- Reading LT04 tape test @ PIC with CMS PhEDEx stager scripts (G03243 – STK)

- [1] & [2] : Complete tape read (278 FILES / 766 GBs)

- **[1]** STKL403 (ts002) - **[2]** STKL405 (ts020)
- Drive rate = **[1]** 83.4 MB/s - **[2]** 80.4 MB/s
- Overall rate = **[1]** 72.6 MB/s - **[2]** 66.5 MB/s
- *f* = drive/overall = **[1]** 0.87 - **[2]** 0.83 → **~15% rate reduction**

- [3] : Alternate file read, based on 'location cookie' → 1,3,5,7... (139 FILES / 382 GBs)

- STKL402 (ts001)
- Drive rate = 81.9 MB/s
- Overall rate = 40.8 MB/s
- *f* = drive/overall = 0.50

- [4] : Alternate file read, with file seeks > 12 GBs in distance (44 FILES / 119 GBs)

- STKL403 (ts002)
- Drive rate = 80.0 MB/s
- Overall rate = 29.2 MB/s
- ***f* = drive/overall = 0.37 → ~60% rate reduction (~ STEP09)**

RAW+RECO files on tapes!
Reading RAW files!

LT04 performance tests conclusions

- **Incomplete tape reads has a negative impact on reading performance…**

- **Complete LT04 tape reads is a must:**

  (a) better file orders via FFs (i.e. better organisation of files in tape groups) **[VOs]**
  (b) pre-stage big portions of data → more chance of reading complete tapes **[VOs]**
  (c) if a VO wants more than xx% of data from a Tape, Enstore could fire a complete tape read (?)… **[SITES]**
  (d) install more LT04 drives (brute force, but effective) **[SITES]**

- Setup sites to pre-stage big portions of data:

  - Increasing MSS recall queue lengths **[SITES]**
  - During STEP'09, at PIC it was increased from 2k to 5k petitions (all VOs)

- Optimize write access to Tape to minimise collisions with read accesses… **[SITES]**

CMS: Improving Reads

- Better organisation of files in tape groups (FFs). At PIC we follow this convention today:
  - **REAL DATA:** create one FF per Primary DataSet + Data Tier (RAW/RECO/AOD)
  - **MC:** as there are a large number of datasets, create 1 FF for different Data Tier

-MC-

t1d0mc_RAW
t1d0mc_RECO
t1d0mc_AOD
t1d0mc_ALCARECO
...

-DATA-

t1d0data_XXX_RAW
t1d0data_XXX_RECO
t1d0data_YYY_RAW
...

- CMS announces LFNs to Tier-1 sites in advance, so they can setup File Families
  - At PIC this FF creation procedure is now automatized

PIC: Improving Writes

- Brand new implemented **pool buffer** to write chunks of files to tape at PIC (not individual files on demand):

  - This aliviates drive occupancy, hence overall MSS system performance

- Writing policy setup per pool:

  >1000 files to migrate **OR** >20 GBs **OR** >2h since buffer was created

- Each FF has an own buffer on the pool

- Enstore **DISCIPLINE** needs to be enabled, so Enstore can process multiple pool migration buffers

  - FNAL developed it, PIC is deploying/testing it atm

PIC: Perfomance tests

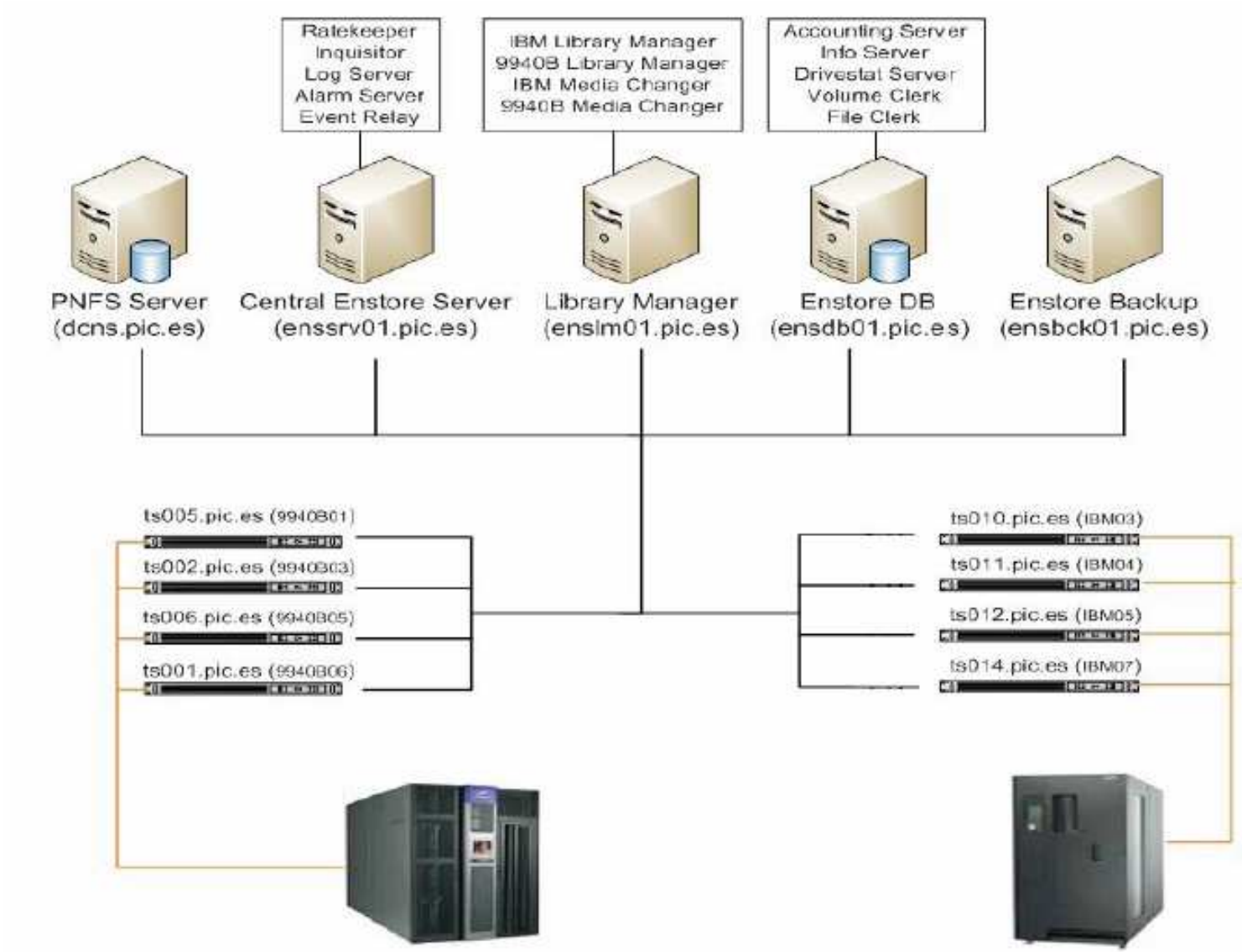
- We plan a custom multi-VO test at PIC to check MSS performance with these new CMS and PIC implementations (once DISCIPLINE is deployed at PIC and before Data Taking)

Conclusions

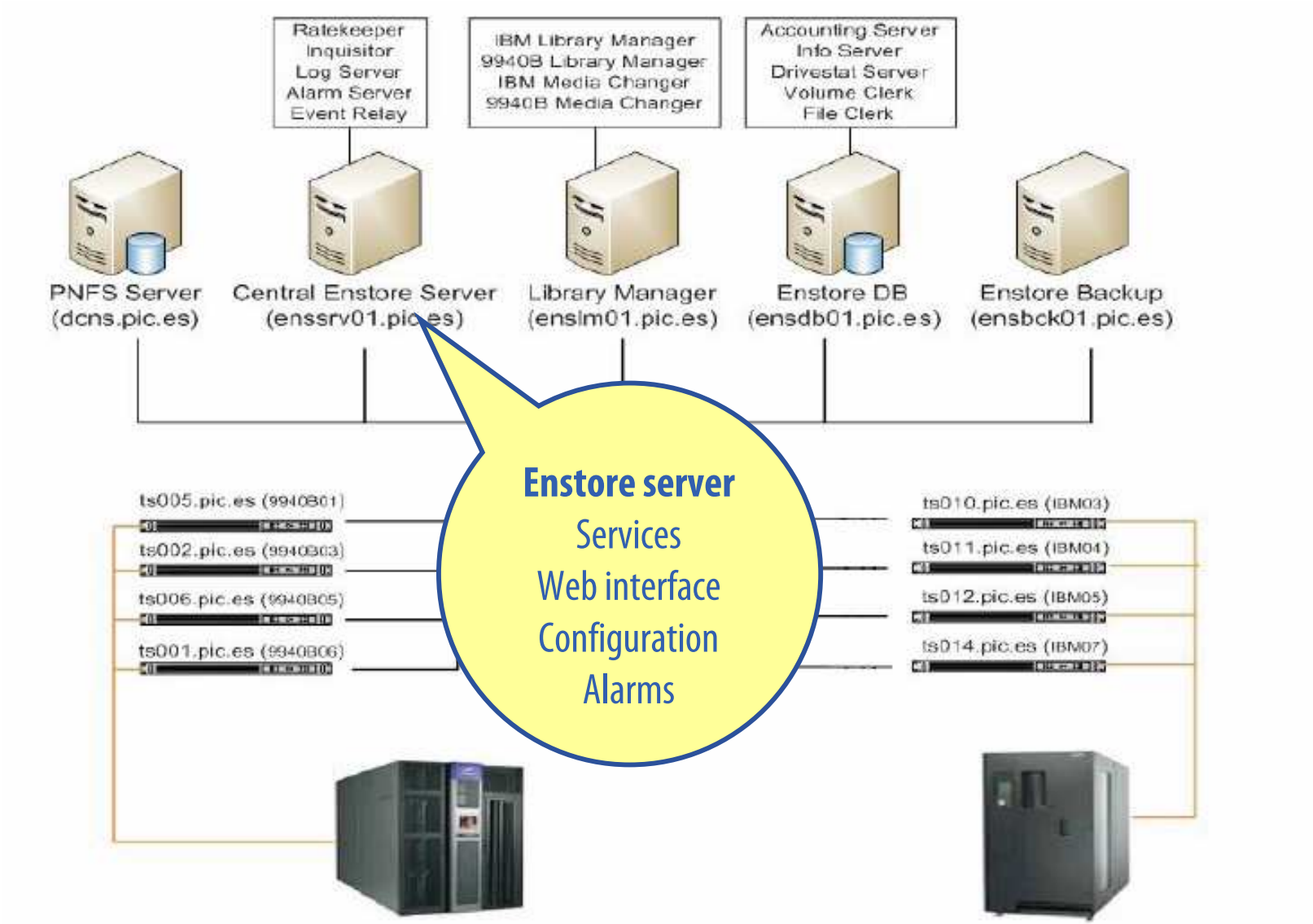
- **Low** read/write throughputs from/to LT04 tapes observed during STEP'09
- To get MSS high read performance, **reading complete tapes is a must!**
  - **CMS:** Better organisation of files in tape groups (FFs) + pre-stage big portions of data
  - **PIC:** Increasing MSS recall queue lengths + install more LT04 drives
    + Optimize write access to Tape to unload drive occupancies
    + increasing recall pools + optimizing pool costs
- **PIC:** Implementation of write buffers to MSS on disk pools (DISCIPLINE)
- CMS plans to have all custodial data at Tier-1s on disk for the 1st year, but has recently established a **working group** to integrate pre-staging on the Workload Management tools
- Once PIC has deployed **DISCIPLINE**, some local tests with all LHC VOs to be coordinated

Backup slides

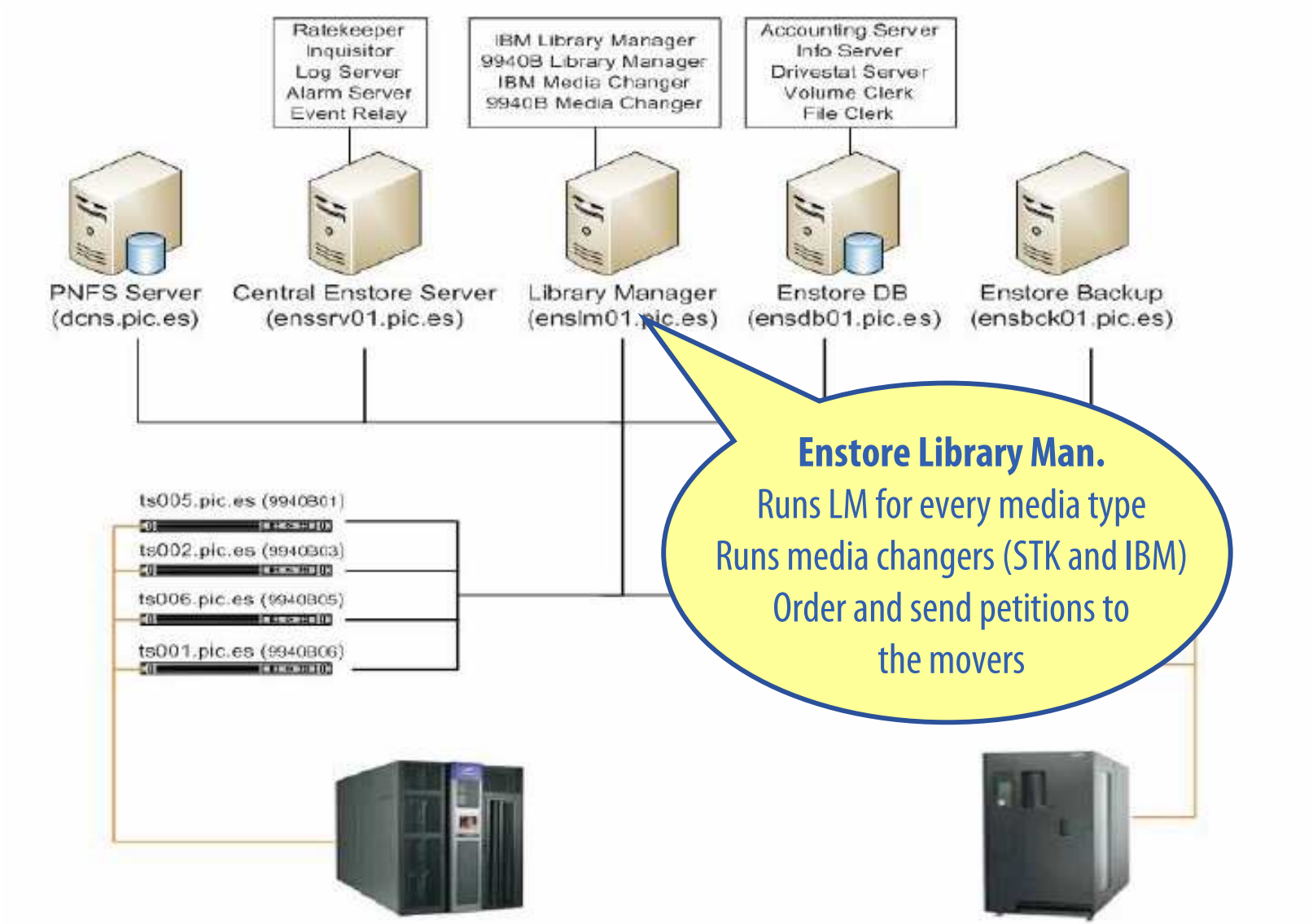
Enstore


Ratekeeper
Inquisitor
Log Server
Alarm Server
Event Relay
IBM Library Manager
9940B Library Manager
IBM Media Changer
9940B Media Changer
Accounting Server
Info Server
Drivestat Server
Volume Clerk
File Clerk
PNFS Server
(dcns.pic.es)
Central Enstore Server
(enssrv01.pic.es)
Library Manager
(ensim01.pic.es)
Enstore DB
(ensdb01.pic.es)
Enstore Backup
(ensbck01.pic.es)
ts005.pic.es (9940B01)
ts002.pic.es (9940B03)
ts006.pic.es (9940B05)
ts001.pic.es (9940B06)
ts010.pic.es (IBM03)
ts011.pic.es (IBM04)
ts012.pic.es (IBM05)
ts014.pic.es (IBM07)

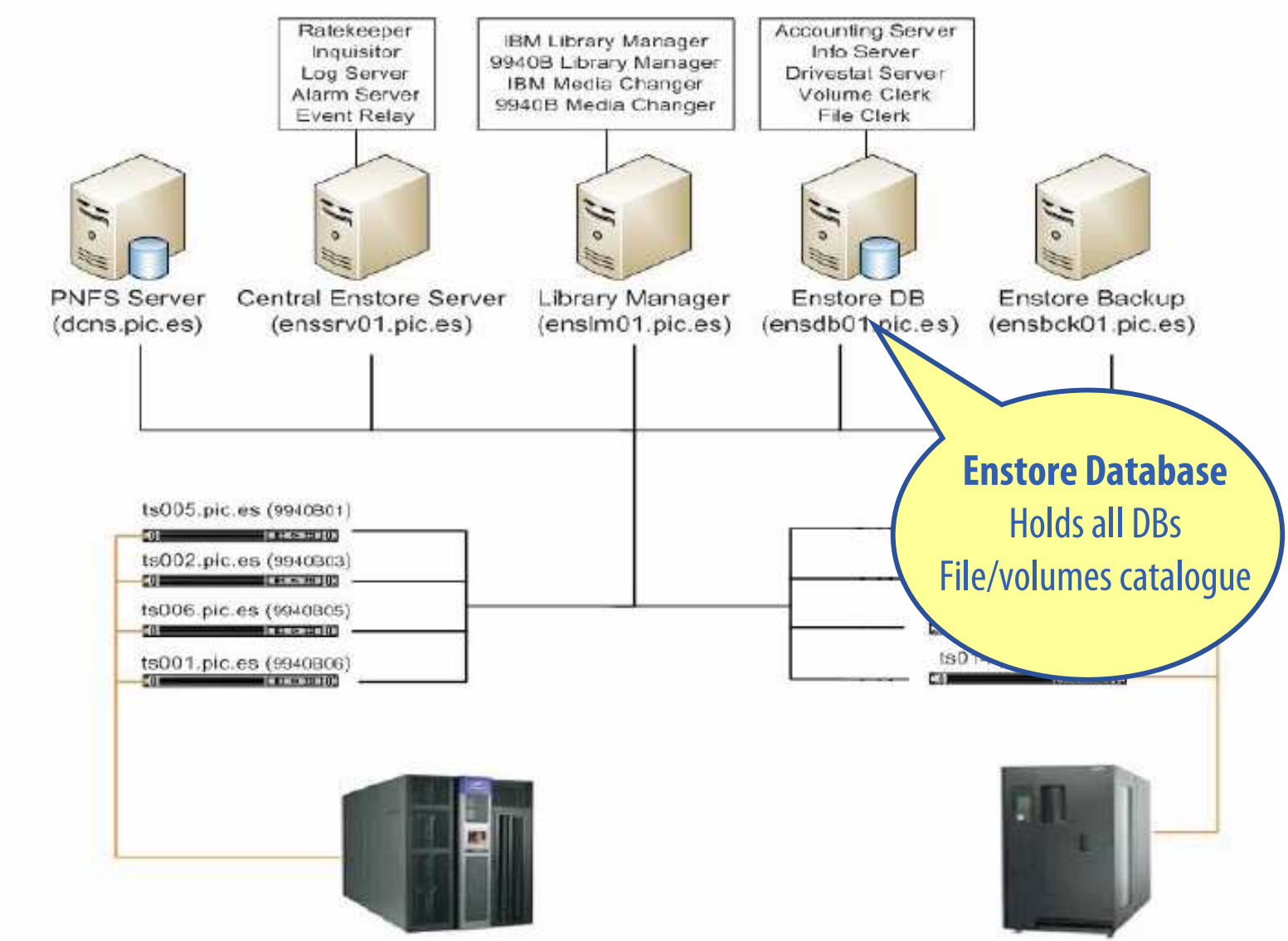
Enstore


Ratekeeper
Inquisitor
Log Server
Alarm Server
Event Relay
IBM Library Manager
9940B Library Manager
IBM Media Changer
9940B Media Changer
Accounting Server
Info Server
Drivestat Server
Volume Clerk
File Clerk
PNFS Server
(dcns.pic.es)
Central Enstore Server
(enssrv01.pic.es)
Library Manager
(ensim01.pic.es)
Enstore DB
(ensdb01.pic.es)
Enstore Backup
(ensbck01.pic.es)
Enstore server
Services
Web interface
Configuration
Alarms
ts005.pic.es (9940B01)
ts002.pic.es (9940B03)
ts006.pic.es (9940B05)
ts001.pic.es (9940B06)
ts010.pic.es (IBM03)
ts011.pic.es (IBM04)
ts012.pic.es (IBM05)
ts014.pic.es (IBM07)

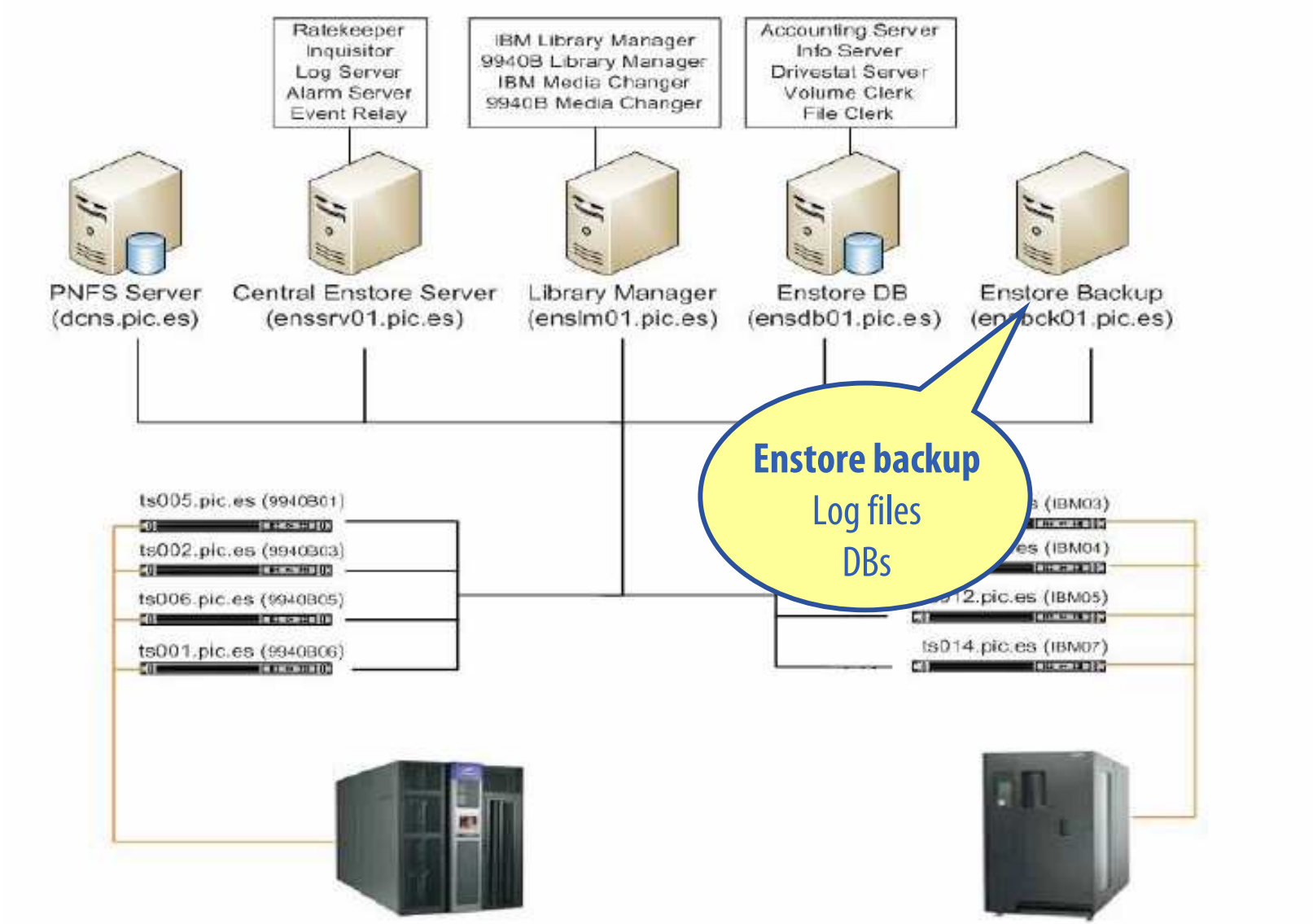
Enstore


Ratekeeper
Inquisitor
Log Server
Alarm Server
Event Relay
IBM Library Manager
9940B Library Manager
IBM Media Changer
9940B Media Changer
Accounting Server
Info Server
Drivestat Server
Volume Clerk
File Clerk
PNFS Server
(dcns.pic.es)
Central Enstore Server
(enssrv01.pic.es)
Library Manager
(ensim01.pic.es)
Enstore DB
(ensdb01.pic.es)
Enstore Backup
(ensbck01.pic.es)
ts005.pic.es (9940B01)
ts002.pic.es (9940B03)
ts006.pic.es (9940B05)
ts001.pic.es (9940B06)
Enstore Library Man.
Runs LM for every media type
Runs media changers (STK and IBM)
Order and send petitions to
the movers

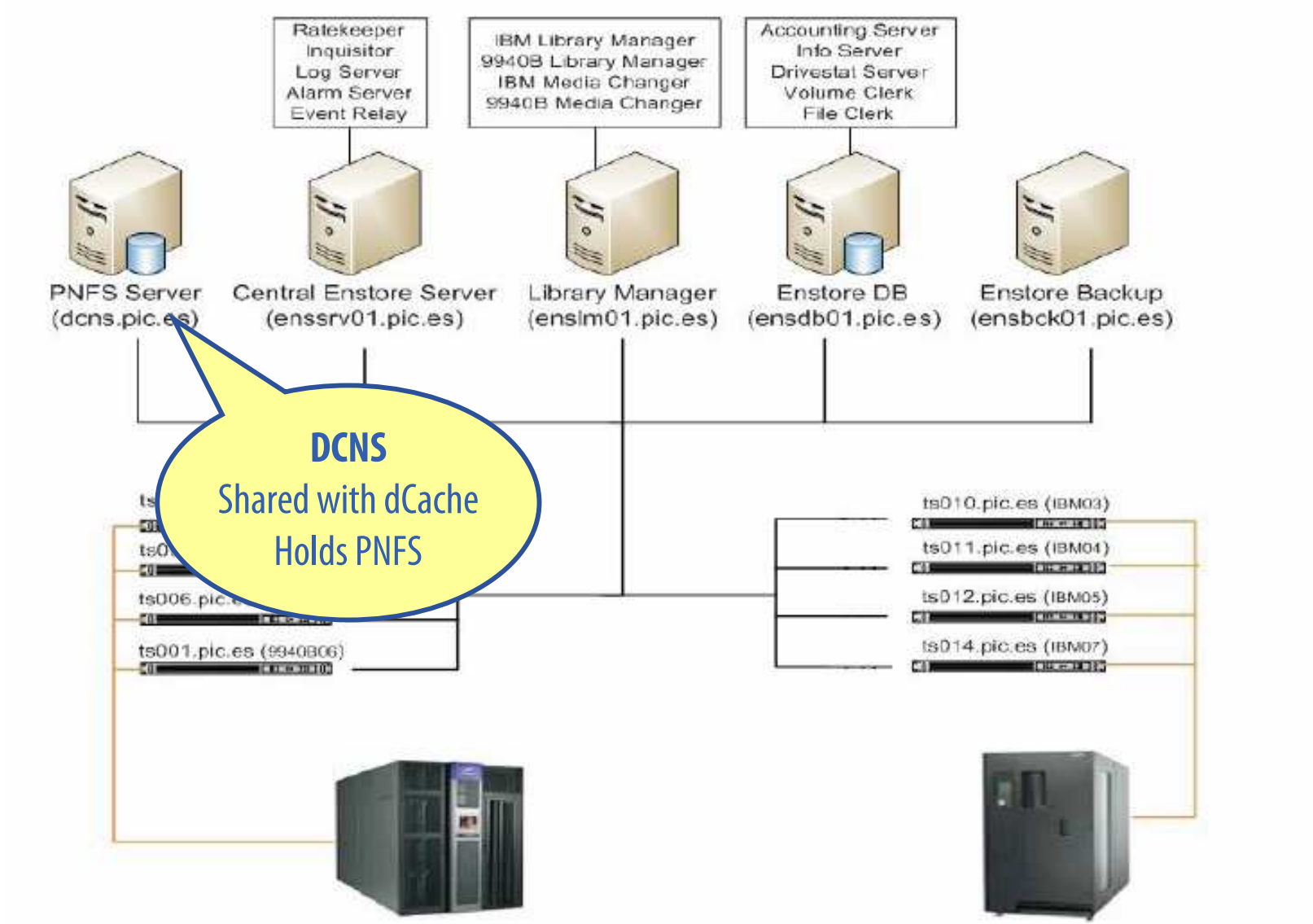
Enstore


Ratekeeper
Inquisitor
Log Server
Alarm Server
Event Relay
IBM Library Manager
9940B Library Manager
IBM Media Changer
9940B Media Changer
Accounting Server
Info Server
Drivestat Server
Volume Clerk
File Clerk
PNFS Server
(dcns.pic.es)
Central Enstore Server
(enssrv01.pic.es)
Library Manager
(ensim01.pic.es)
Enstore DB
(ensdb01.pic.es)
Enstore Backup
(ensbck01.pic.es)
ts005.pic.es (9940B01)
ts002.pic.es (9940B03)
ts006.pic.es (9940B05)
ts001.pic.es (9940B06)
Enstore Database
Holds all DBs
File/volumes catalogue

Enstore


Ratekeeper
Inquisitor
Log Server
Alarm Server
Event Relay
IBM Library Manager
9940B Library Manager
IBM Media Changer
9940B Media Changer
Accounting Server
Info Server
Drivestat Server
Volume Clerk
File Clerk
PNFS Server
(dcns.pic.es)
Central Enstore Server
(enssrv01.pic.es)
Library Manager
(ensim01.pic.es)
Enstore DB
(ensdb01.pic.es)
Enstore Backup
(ensbck01.pic.es)
Enstore backup
Log files
DBs
ts005.pic.es (9940B01)
ts002.pic.es (9940B03)
ts006.pic.es (9940B05)
ts001.pic.es (9940B06)
ts014.pic.es (IBM07)

Enstore


Ratekeeper
Inquisitor
Log Server
Alarm Server
Event Relay
IBM Library Manager
9940B Library Manager
IBM Media Changer
9940B Media Changer
Accounting Server
Info Server
Drivestat Server
Volume Clerk
File Clerk
PNFS Server
(dcns.pic.es)
Central Enstore Server
(enssrv01.pic.es)
Library Manager
(ensim01.pic.es)
Enstore DB
(ensdb01.pic.es)
Enstore Backup
(ensbck01.pic.es)
DCNS
Shared with dCache
Holds PNFS
ts001.pic.es (9940B06)
ts010.pic.es (IBM03)
ts011.pic.es (IBM04)
ts012.pic.es (IBM05)
ts014.pic.es (IBM07)