IN2P3 site report
HPSS User Forum 2013

Pierre-Emmanuel BRINETTE
Benoit DELAUNAY

November 5, 2013

Agenda

- IN2P3
- HPSS usage
  - RFIO
  - Treqs
- Infrastructure
- Monitoring
- Issues and wishes

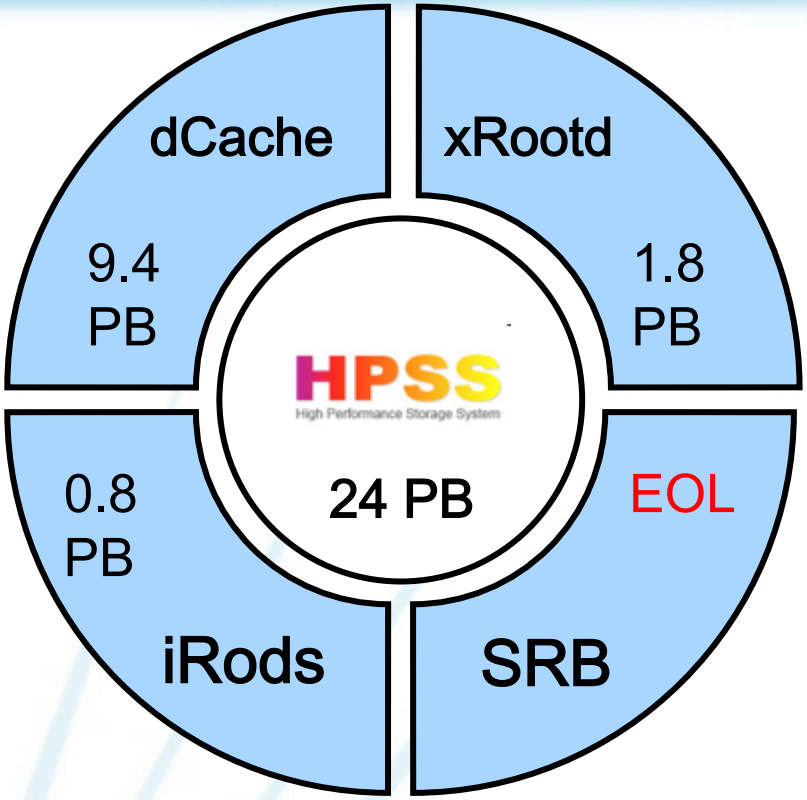
IN2P3, who are we ?

- IN2P3, is the French National Institute for Nuclear and Particle Physics Research.
- Composed of 20 laboratories, 3 experimental sites and 1 computing center, all distributed across the French territory.
- Around 2500 collaborators, including 900 researchers.
- Involved into international experiments of High Energy Physics and Astrophysics
  - LHC
  - LSST / Planck / AMS

CCIN2P3, who are we ?

- IN2P3 Computing Center (Lyon)
- 2 Computing rooms (800 m² each)
- Resources :
  - Computing cluster (DELL):
    - ~ 18k cores
    - 190 k HEP-Spec06 / hours
  - Disk storage
  - MSS
  - End user services : mail / backup / web / DB / $HOME …
- 80 employees
- 2000 active users
- EGI / WLCG Grid
- HPSS since 1999

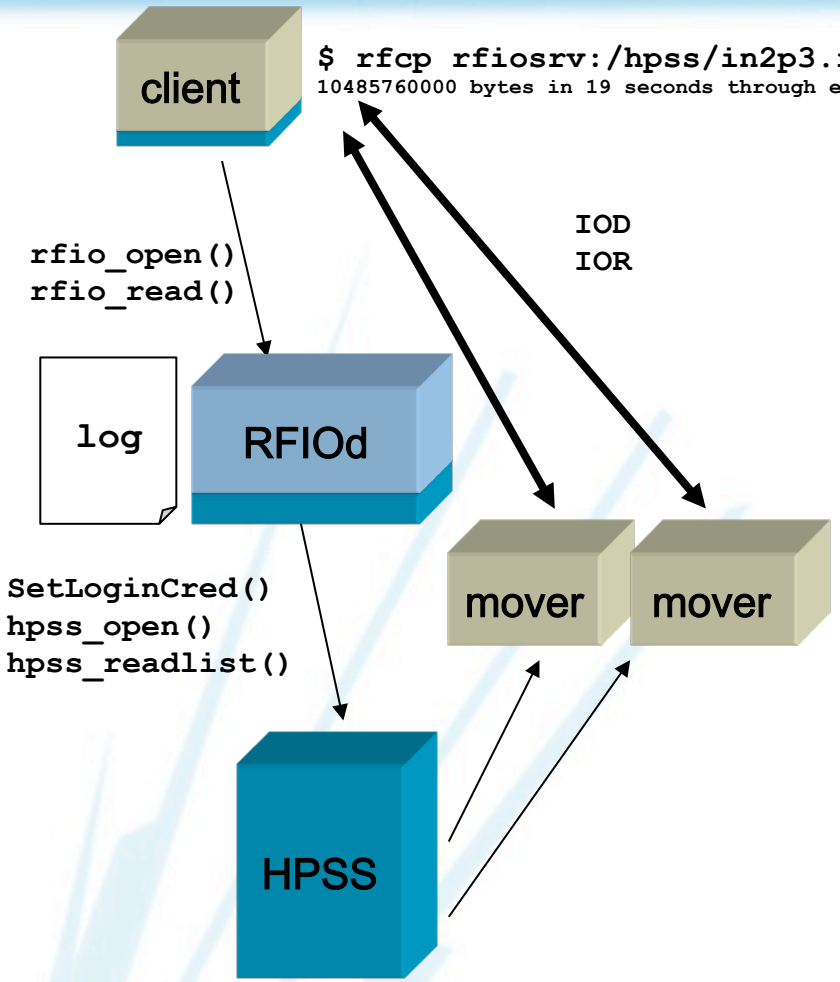
Storage @ IN2P3



Cluster Filesystem
1.4 PB GPFS 3.5

- Half of data managed by external storage software
  - dCache + xrootd : LHC
  - iRods
- Large disk capacity (~ 12 PB) using DAS server :
  - Sun X4540 (130)
  - DELL R510 (110)
- GPFS : /scratch
- HPSS :
  - Still direct user access
  - More and more used as tape backend
  - Grow + 5 PB / years

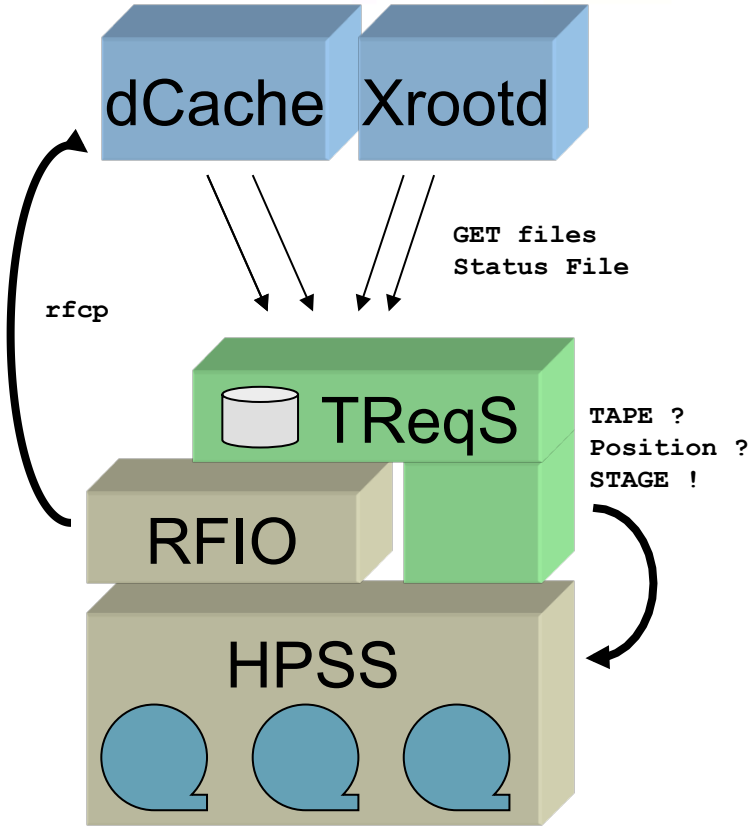
HPSS data access



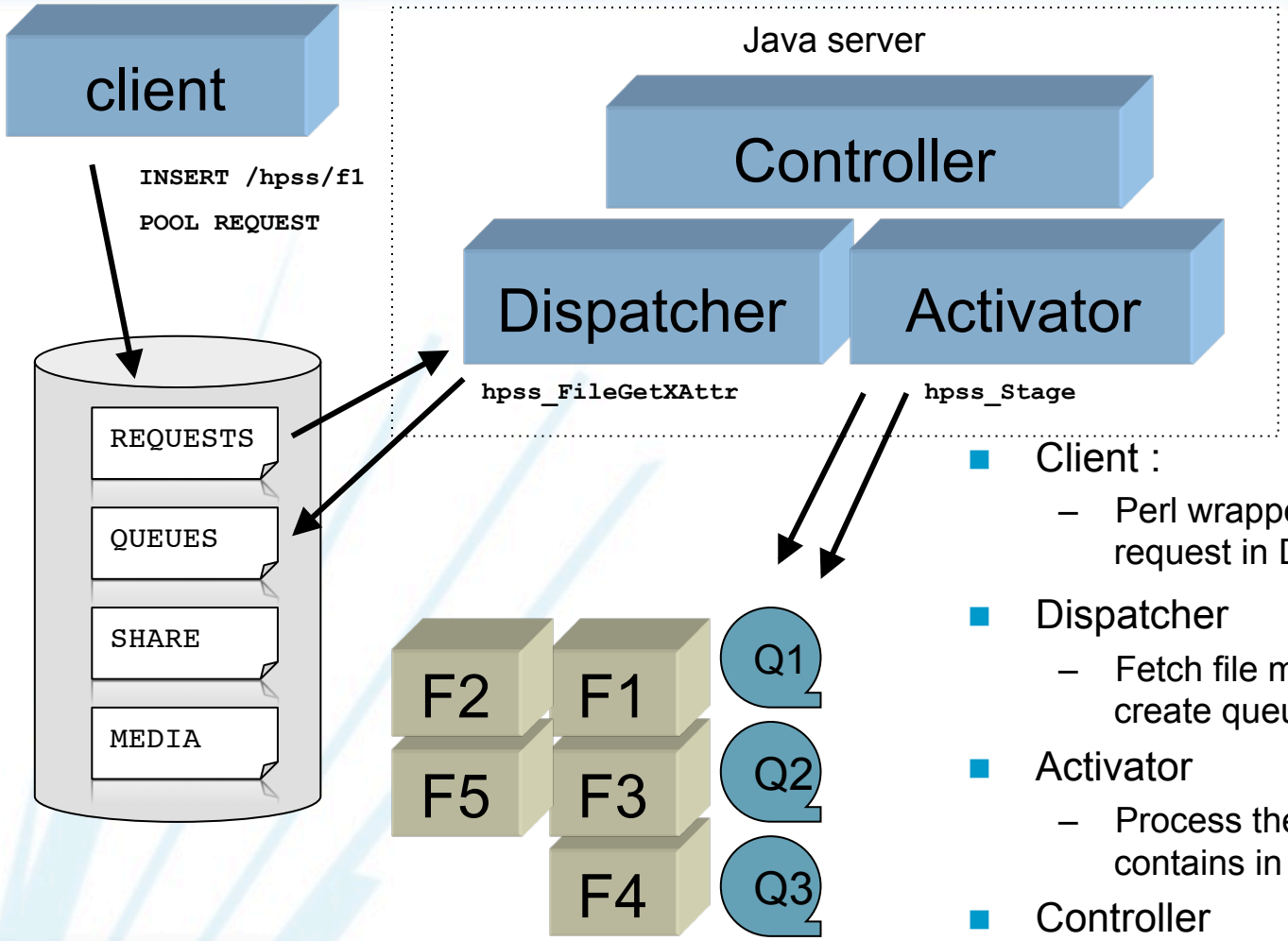
- RFIO
  - Provide Unix like command to end users : rfcp , rfdir, rfrm, …
  - rfcp client use HPSS API
    - `hpss_readlist`
    - `hpss_writelist`
- Benefits
  - Good performances, direct transfers from movers
  - hpss libraries statically linked on the client
  - Access trough a control server
    - Limit simultaneous Cx
    - Logging

TReqS

- TReqS : Tape Requests Scheduler
  - Idea stolen from BNL Batch (ERADAT)
  - Thanks to David Yu
- HPSS Frontend for optimizing staging (read) operations
  - Sort requests by TAPE and Position
  - Limit tape mounting / Dismounting
  - Increase bandwidth for read operations when data are spread over many tapes
  - Optimize tape drive usage
  - Reduce the impact of massive staging against HPSS
- Handle all read operations for dCache, xRootd and iRODS



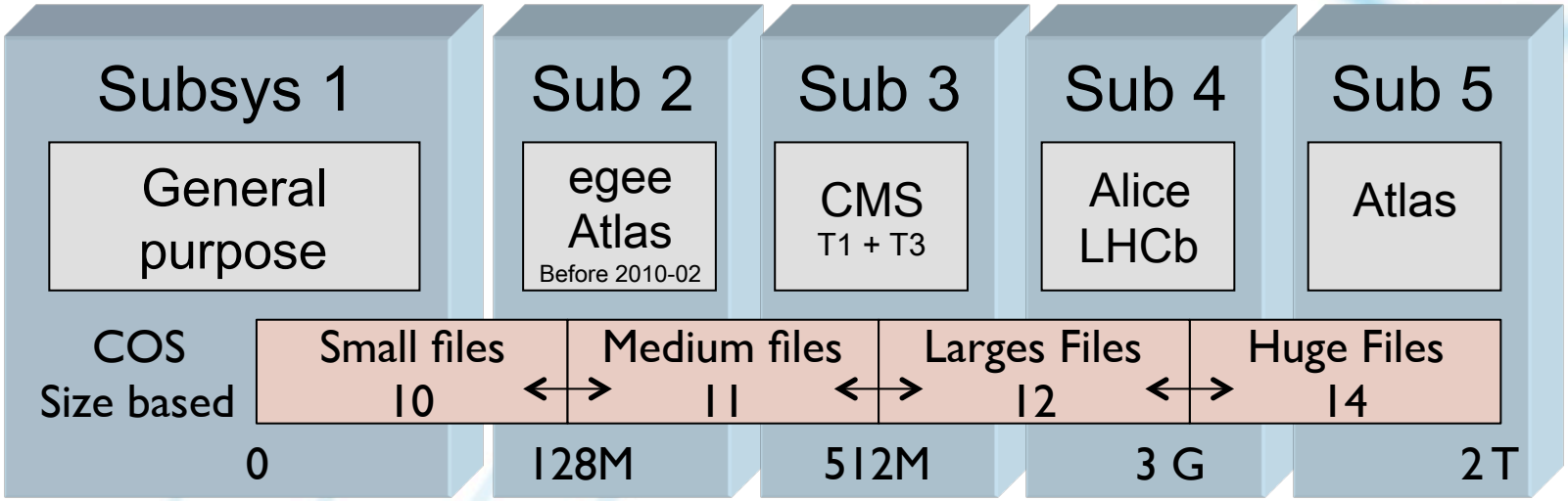
TReqS



- Client :
  - Perl wrapper that insert / pool file request in DB
- Dispatcher
  - Fetch file metadata from HPSS and create queue
- Activator
  - Process the queues : Stage the files contains in a queue
- Controller
  - Trigger Dispatcher and activator the according the share policy

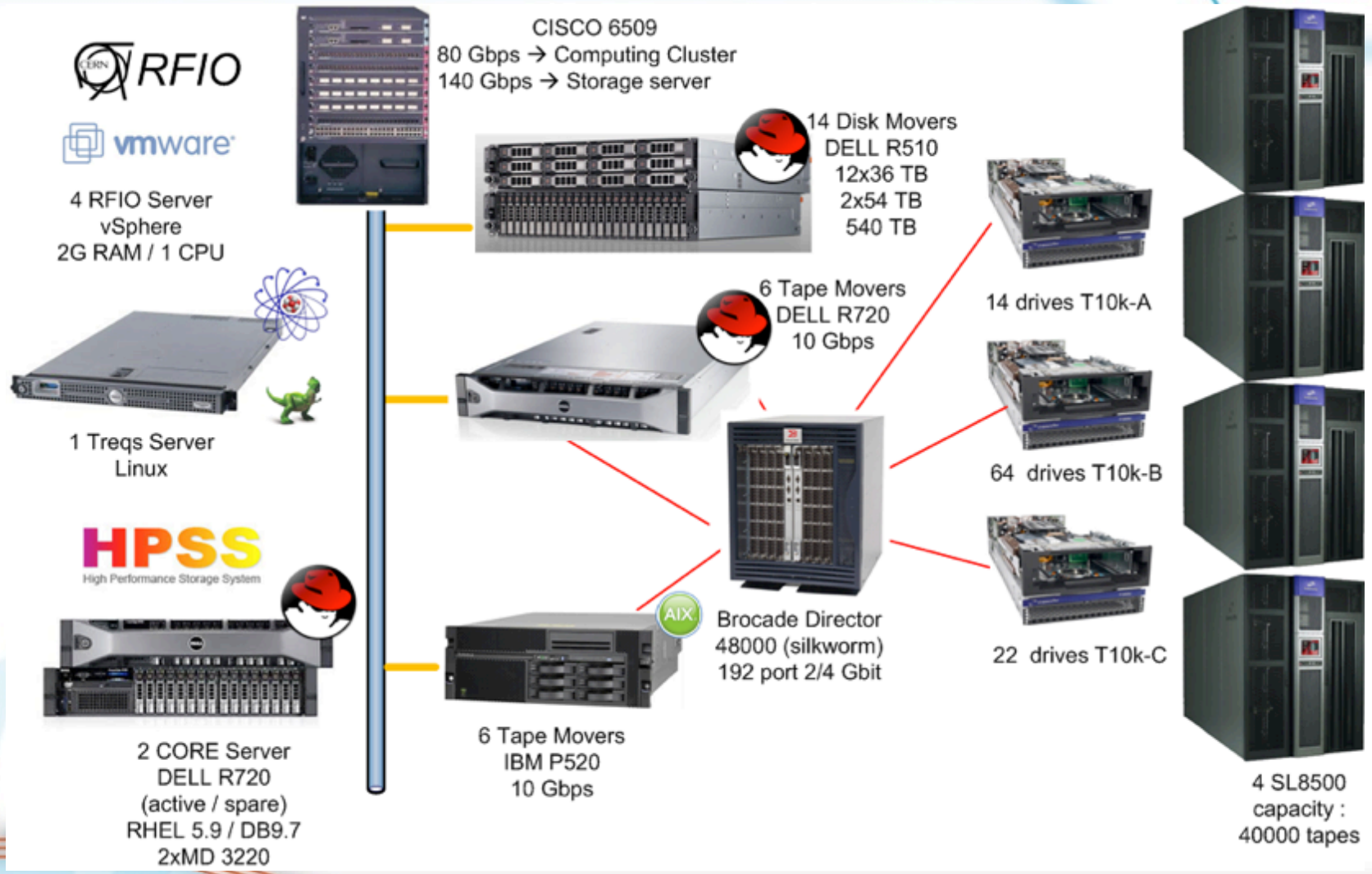
Storage policy

HUF 2012



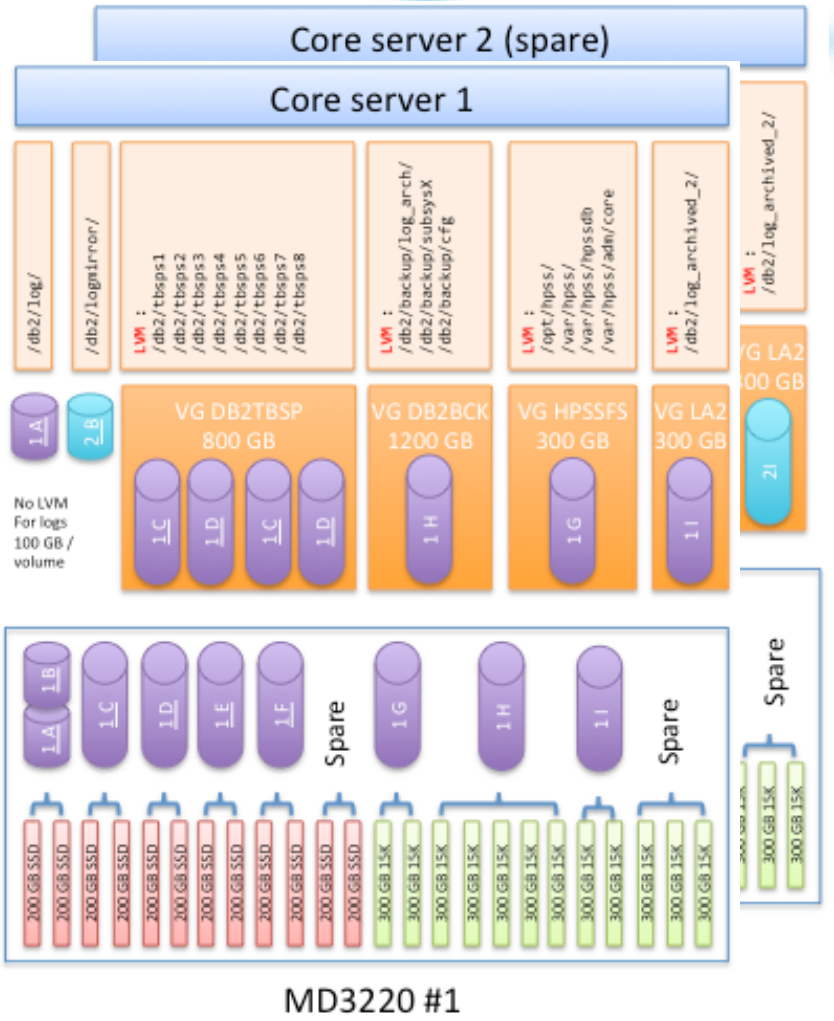
- Historical
  - 11.5 PB
  - ~2000 UID
  - 33 M files
- Newly created
  - 11.5 PB
  - 12 M files
  - Mainly used for LHC
- Dedicated subsystem
  - Allow to dedicate DISK resources for specific set of users when using automatic COS selection
  - Specific database for a set users → faster query
    - Subsys 1 : 40 GB
    - Subsys [2-5] : 1.5 to 6 GB

Infrastructure



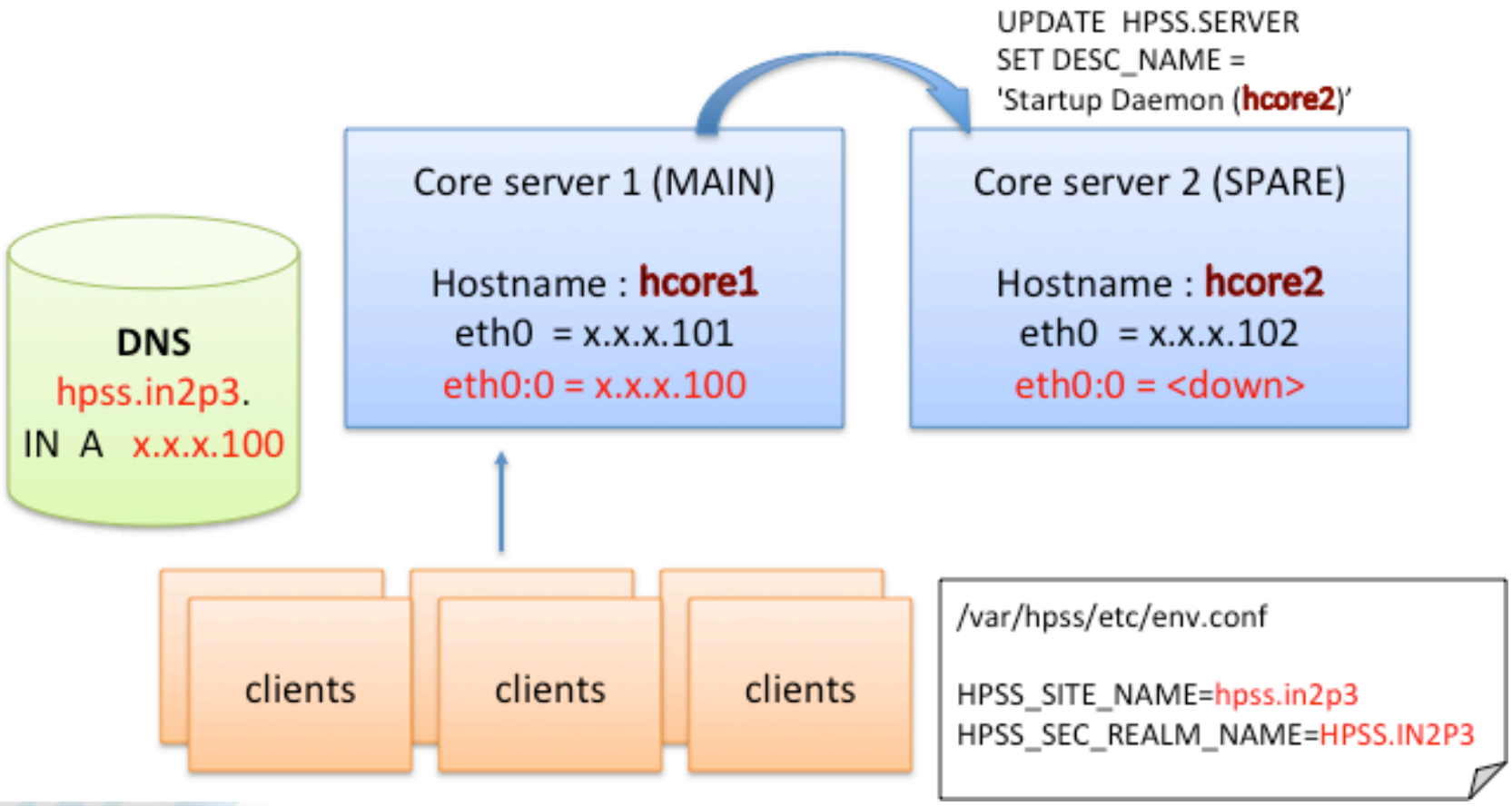
Core servers

- 2 x DELL R720 (Active / spare)
  - 2 x 8 core HT (Xeon E5-2690)
  - 128 GB RAM
- 2 x DELL MD3220
  - 12 SSD + 12 SAS 15K
- DB2 Metadata :
  - Log on 1 x RAID 1 SSD
  - Mirror log on 1 x RAID 1 SSD (2nd unit)
  - Tablespaces on 4 x RAID 1 SSD
  - Backup + archived log on SAS RAID 5
- RHEL 5, HPSS 7.3.3p8, DB2 9.7fp8
  - AUTOMATIC STORAGE



Core servers

- "REALM_NAME" associated to a "floating" IP



Tape Storage

- 4 SL8500
  - 40 k slots
  - T10K-A : 12 drives (phase out)
  - T10K-B : 64 drives
  - T10K-C : 22 drives
  - 4000 mounts / day
- Media :
  - T10K-A : 2 000
  - T10K-B : 21 000
  - T10K-C : 1 000
- T10K-A “retired” since Q2 2012
  - Repack **still** in progress
  - Media recycled as T10K-B
  - EOL expected in Q1 2014
- T10K-B phase out planned soon
  - + 20.000 Tape T10K-T1 (A and B) to repack
  - ~ 80.000 estimated hours of repacks
  - Almost 10 years with a single drive !

- Library Management
  - ACSLS v 8.2
  - Monitoring using Storesentry
  - Looking on Oracle Tape Analytics

Tape Storage

- Media warranty information in ACSLS 8.2

```
display volume * -f  entry_date access_count
end_of_life warranty_life load_limit_alert -s
end_of_life
```

```
ACSSA> display volume JS0* -f  entry_date access_count end_of_life warranty_life load_limit_alert -s
end_of_life
2013-11-05 16:29:42                Display Volume
 Vol_id              Entry_date           Access_count  End_of_life  Warranty_life  Load_limit_alert
[…]
 JS0353              2012-08-09 15:00:31  129           38.4%         57.5%         before_load_limit
 JS0377              2012-08-09 15:00:28  1279          67.7%        101.5%         past_load_limit
 JS0369              2012-08-09 15:02:09  1849          75.5%        113.3%         before_load_limit
```

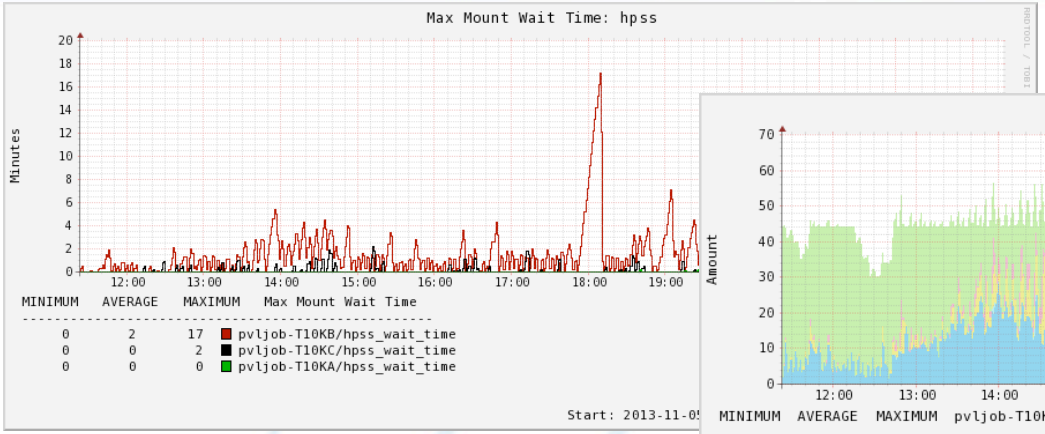
Tape Storage

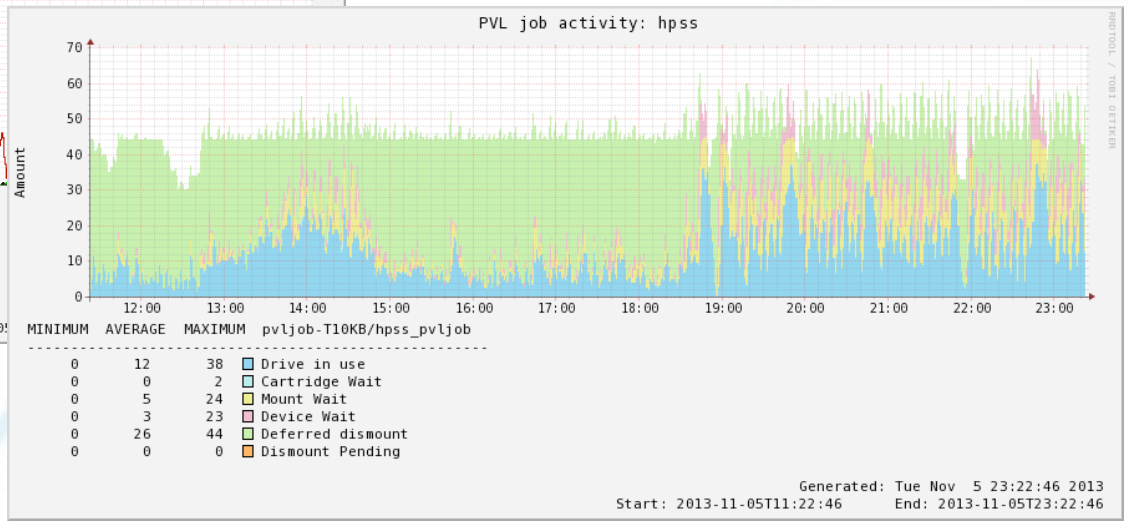
- Tape mover :
  - "Old" p505/p515 retired after 8 years in production.
  - 6 Linux R720
    - 2 HBA
    - 10 Gbps
    - Up to 15 drives per mover (mix A/B)
  - 6 AIX p520 (2 HBA + NIC 10 Gbps)

Disk Storage

- Since 2011 : DAS
  - 12 Dell R510 + MD 1200
    - 24 x 2 TiB SAS Nearline Disk
    - 10 Gbps NIC
    - 36 TiB RAID 6
  - 432 TiB for HPSS
- Q2 2013
  - + 2 Dell R510 + MD1200
    - 24 x 3 TiB SAS NearLine
    - 10 Gbps NIC
    - 54 TiB RAID 6
- Total : 540 TB

Monitoring





- Mount activity monitored using collectd
  - DB2 query every minutes over HCFG.PVLxxx
  - Usefull to monitor drives load detect drive or tape issues

Monitoring

Service	Status
Check hpss mount time for T10KA	OK
Check hpss mount time for T10KB	OK
Check hpss mount time for T10KC	OK
Check storage class	OK

- Nagios monitoring
  - Alarm if a mount request waits more than 20 minutes.
    - Data grabed from collectd
  - Alarm if storage class status goes Warning or Critical
    - Hpssadm script every 2 hours.

Monitoring

ACS	ERROR	2013-11-05 18:04:26 Emsg 'cps_getline: st_can_current() unexpected status = STATUS_QUEUE_FAILURE'
ACS	ERROR	2013-11-05 18:04:26 Emsg 'cps_getline: cl_qm_mlocate() unexpected status = STATUS_PROCESS_FAILURE'
IPMIEVT	ERROR	SWATCH:ccsvli69: Fan fault (PS 2)
HPSS	WARNING	MINR LKT10000 RC=0 pbrinett@cchcsli001: Volume JT645100 state changed to DOWN by user pbrinett
HPSS	WARNING	WARN PVRS0141 RC=0 Robot unable to find cartridge, cartridge = JT6451, drive = 0
HPSS	WARNING	WARN PVRS0141 RC=0 Robot unable to find cartridge, cartridge = JT6451, drive = 0
HPSS	WARNING	EVNT MOVR2001 RC=110 Mover Tape (cchmtrs032): End of media on device 208, volume JTG53000, section 257, offset 0.508559360
HPSS	WARNING	EVNT MOVR2001 RC=110 Mover Tape (cchmtrs032): End of media on device 208, volume JS072400, section 5748, offset 0.35651584
DCACHE	WARNING	Check LCG cells : 1 NON-critical cell(s) OFFLINE in lcg
SIMONE	**ALERT**	**alert, "/vms" is over limit (60%)**
HPSS	WARNING	WARN PVRS0259 RC=-8001 Cartridge not readable in drive, will retry in another drive, cartridge = JTC213, drive = 223

- RLS : Central logging system
  - Tim Starrin's whpss + perl script to filter messages

Burning issues

- None this year !
  - Repack issues fixed on 7.3.3p8
  - Workaround for “group” accounting
    - DB2 query over NSOBJECT and BITFILE now run within 2 minutes
- But still interested by :
  - Job Cancellation (CR 280)
  - Freeing stuck I/O (CR 183)
  - Media read flexibility
  - Quota management

Wish list

- Media lifetime management
  - Automatic check tapes that haven't be mounted for years
  - Automatic repack of tapes when exceed lifetime
  - Alternative : integration with third party monitoring tools (STA, Storesentry, RVA …)
- Smart tape scheduler for recall operations
  - Treqs fit our today needs but this feature should be included in HPSS.