Lustre-HSM at CINES

Integration & administration feedback

22/09/2015

Jérôme Chapelle <chapelle@cines.fr>

Hervé Toureille <toureille@cines.fr>

Agenda

- The CINES
- A datacentric architecture : The Evolution
- Current architecture
- Lustre-HSM in real life : Examples
- Conclusion

CINES

The national computer center of french higher education

- French public organisation under the supervision of the French Ministry in charge of higher education and research.
- Provides the french public research community with computing resources and services.
- Located in Montpellier,

55 persons : technicians, engineers and administratives.

CINES

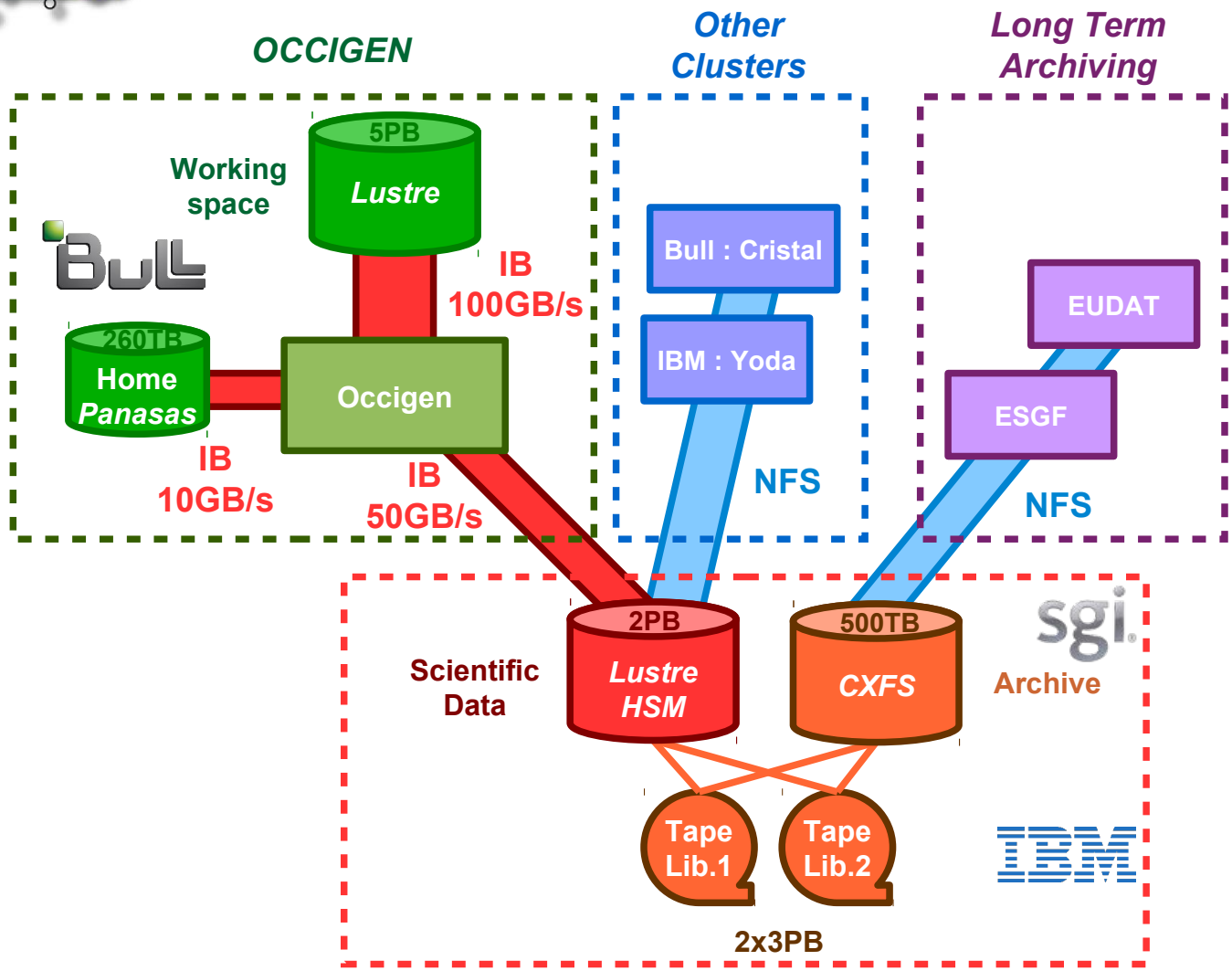
The national computer center of french higher education

Missions :

- High Performance computing : IBM and Bull
- Long term preservation of data and digital documents for universities and public research institutions
- Data center hosting for french national level academic institutions : 10 partnerships, 30 IT cabinets

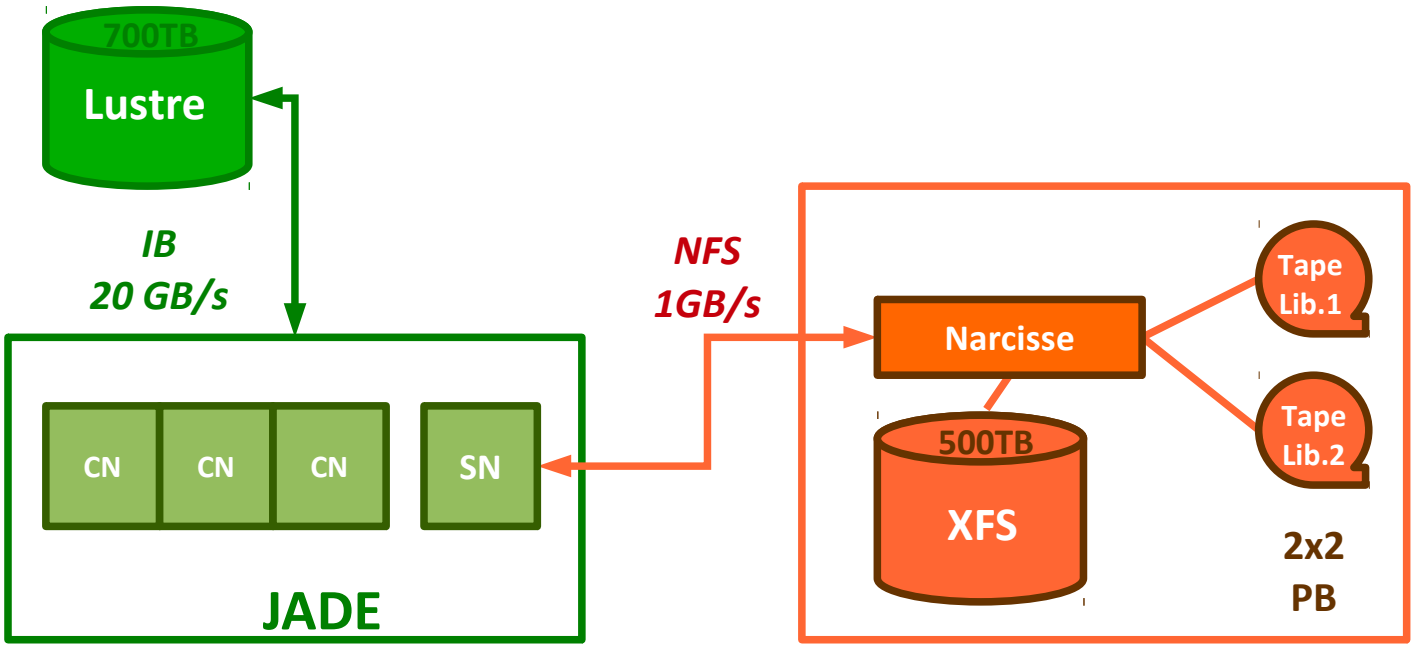
A datacentric architecture

- Why :
  - HPC have their own fast storage, not secured
  - Need to share data between different environments
  - HPC lives and dies, Data remains
- From NFS to Lustre HSM : a « space » odyssey
- Current architecture : machines and processes



Datacentric architecture

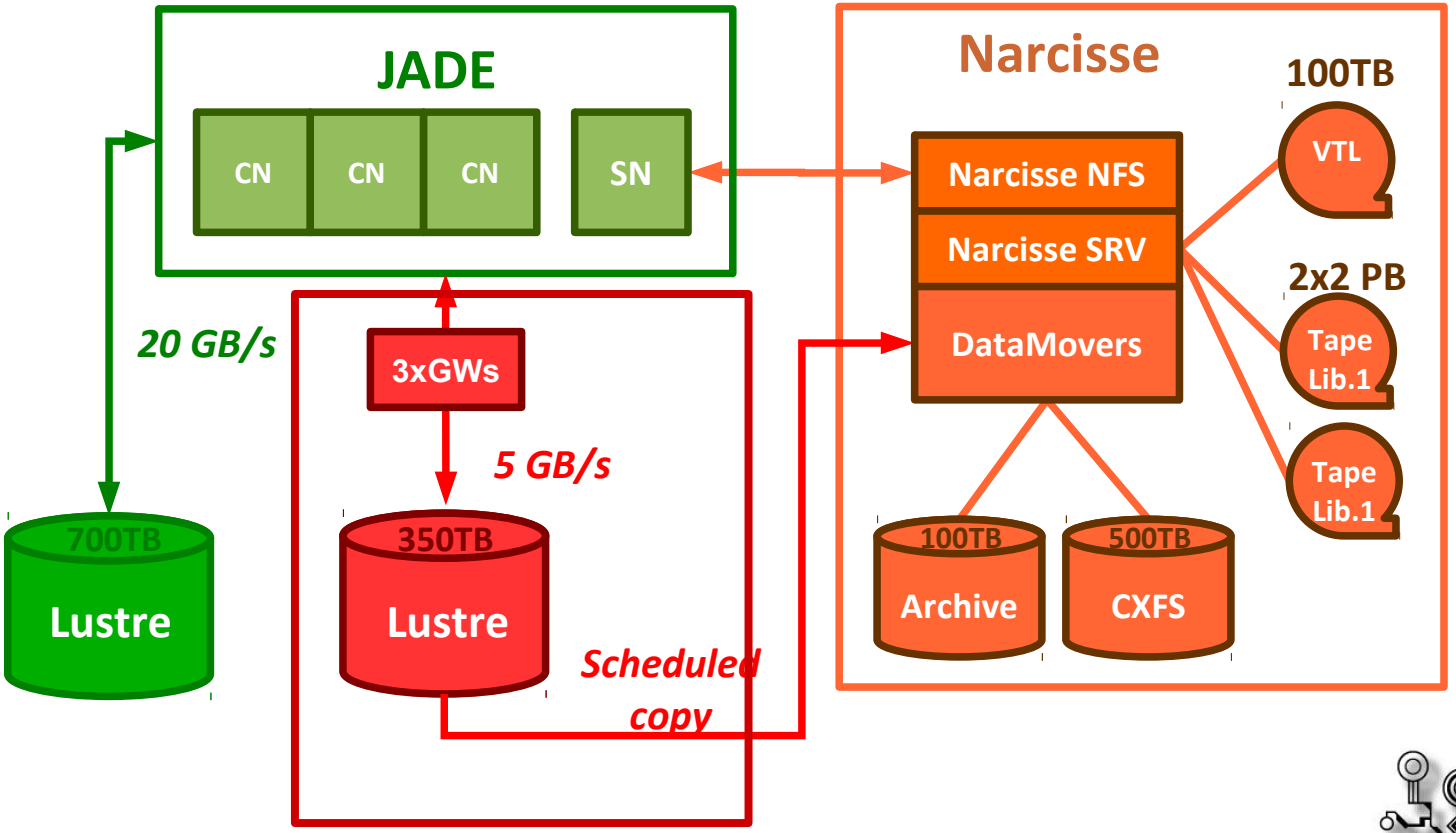
First step - 2008
Narcisse V1



First step - 2008
Narcisse V1 : NFS File server

- 500 TB disks @ 1GB/s (NFS) + 1.5 PB tape
- Offered services :
  - « Unlimitted » storage thanks to DMF HSM
  - Backup and restore service with 15 days retention (based on xfsdump)
- How : DMF, XFS

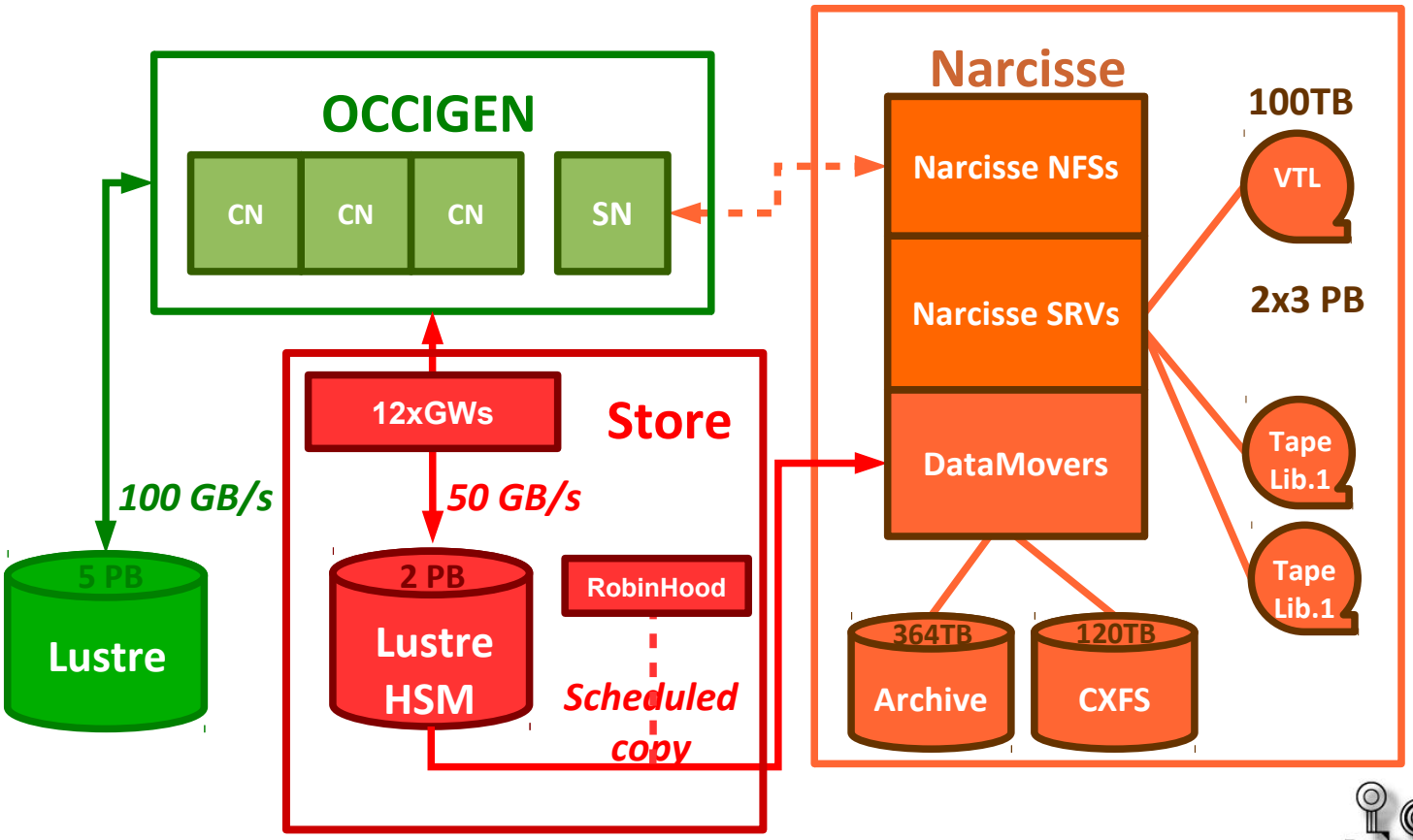
Second Step - 2013



Second Step - 2013
Store 1 : a secured Lustre

- Added 350 TB @ 5GB/s (Lustre)
- Added a 100 TB VTL for small files
- Offered services :
  - Fast and secured storage
  - Compute nodes able to access this space
- How : rsync, pDMF, CXFS
- Expected Lustre HSM, but rsync used instead

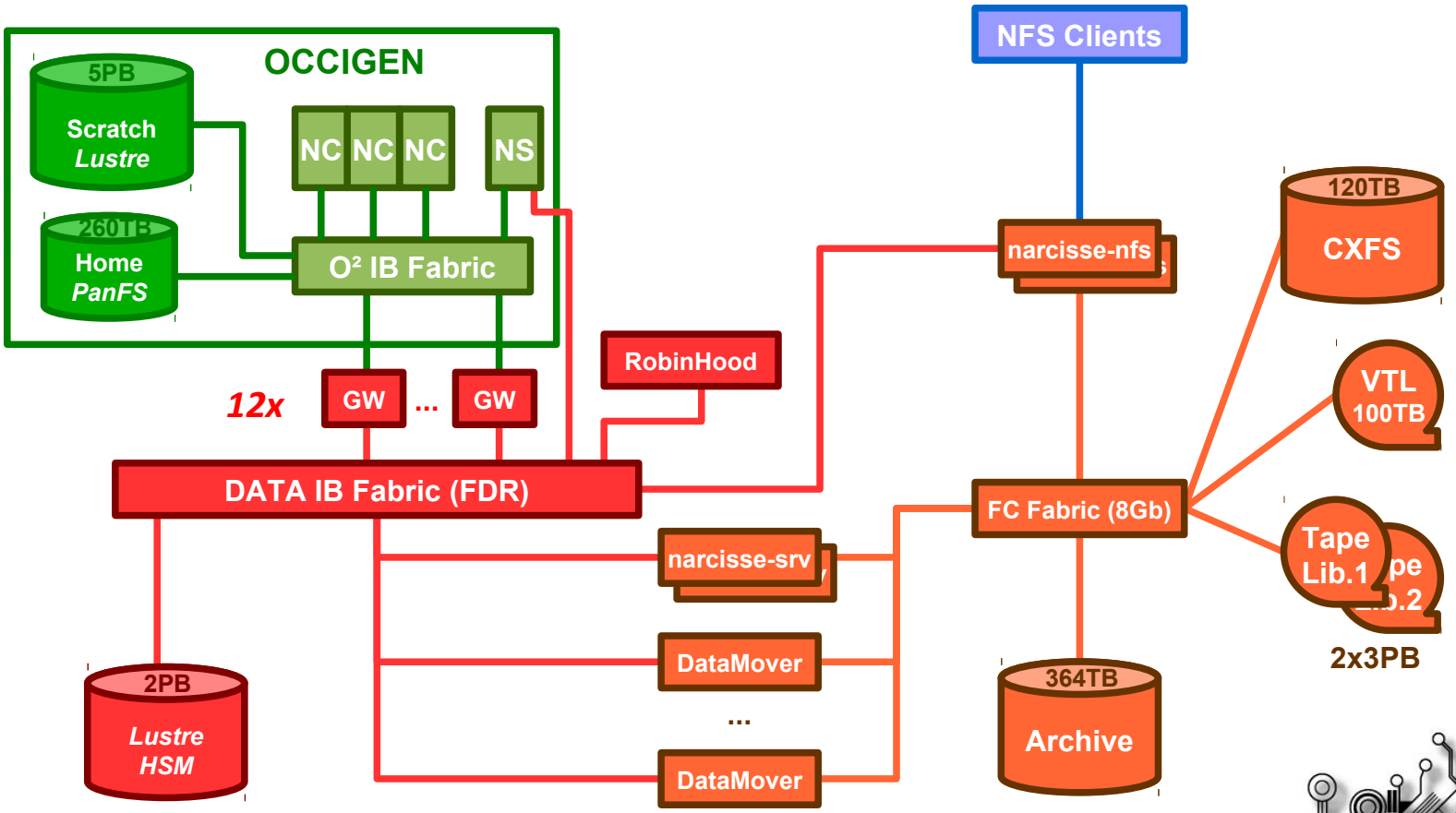
Current Architecture - 2015



Current Architecture - 2015
Store 2 : Lustre HSM

- Upgraded Lustre : 2PB @ 50GB/s with HSM
- High availability cluster
- Offered services :
  - Faster storage
  - « Unlimited » volume thanks to HSM
  - Automated secure process
  - No more easy restore process

Current Architecture - 2015

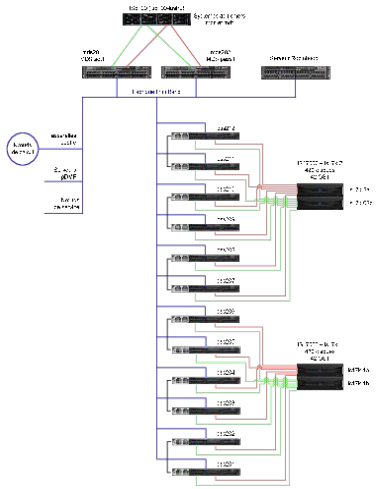


The machines

- Lustre Cluster
- DMF Cluster
- RobinHood

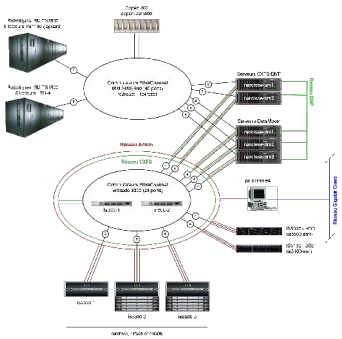
Lustre HSM

- 2x DDN SFA12k
- 420x 3TB SAS Disks
- 2x MDS
- 12x OSS – 84 OST
- 12x LNET Gateways
- InfiniBand FDR (56Gb)
- Lustre 2.5.34.1 (IEEL 2.2.0.2) – RHEL 6



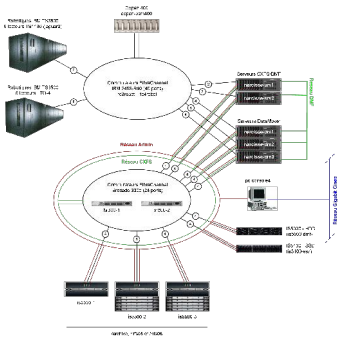
PDMF & CXFS

- 2x CXFS & DMF servers (HA)
- 3x DMF DataMovers
- 2x NFS servers
- 2x IS5500 (NetApp E5400)
- 120x 2TB SAS disks
- 2x Brocade 300e (24x8Gb)

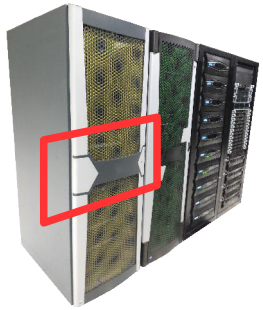


PDMF & CXFS

- 2x CXFS & DMF servers (HA)
- 3x DMF DataMovers
- 2x NFS servers
- 2x IS5500 (NetApp E5400)
- 120x 2TB SAS disks
- 2x Brocade 300e (24x8Gb)



Robinhood



- Intel Xeon E5-2620v2 (2.1Ghz x 6)
- 128GB DDR3
- SSD 460GB : database
- 2x SAS 1TB (RAID1) : system

Tape Libraries

- Two IBM TS3500, with 3000 tapes each (~3PB)
  - Main one : 9 IBM 3592E06 drives
  - Second one : 10 LTO4 drives
- Drive performances : 150 MB/s
- Low TB cost + extensibility

Virtual Tape Library

- One COPAN 400 shelf (100 TB)
- Massive array of idle disk : low energy footprint
- Fast random access (compared to tapes)
- Perfect for small files (90 % of files are small)
- Discontinued product

Lustre in real life

What can we do with a Lutre-hsm pDMF cluster ?

- Securing data
- Gaining space on the Lustre filesystem
- Restoring erased files

Securing Data

```
narcisse-nfs2:/store/lad15_demo # ls -l
-rw-r--r-- 1 demo demo 1447447676 14 sept. 15:43 file1G

narcisse-nfs2:/store/lad15_demo # lfs hsm_state ./file1G
./file1G: (0x00000000)
narcisse-nfs2:/store/lad15_demo # lfs hsm_archive ./file1G
narcisse-nfs2:/store/lad15_demo # lfs hsm_state ./file1G
./file1G: (0x00000000)
```

Nothing happens ?

Beware, migration is an ansynchronous process !

MDS /proc file : active requests

```
[root@mds201 hsm]# cat /proc/fs/lustre/mdt/store2-MDT0000/hsm/active_requests
fid=[0x200010ccf:0x2c08:0x0] dfid=[0x200010ccf:0x2c08:0x0]
compound/cookie=0x5660b019/0x5605081f action=ARCHIVE archive#=1 flags=0x0 extent=0x0-
0xffffffffffffffff gid=0x0 data=[] canceled=0 uuid=9544894b-ea3c-4113-02a9-e8fbcd84b672 done=0
```

Action= ARCHIVE
Fid=[0x200010ccf:0x2c08:0x0]

```
[root@mds201 hsm]# lfs  fid2file /store [0x200010ccf:0x2c08:0x0]
/store/CINES/cnu0003/toureille/lad15_demo/file1G
```

DMF copytool : Logs

```
2015/09/14-16:01:20.507714 lhsmtool_dmf[21725]: copytool fs=store2 archive#=1 item_count=1
2015/09/14-16:01:20.507771 lhsmtool_dmf[21725]: waiting for message from Lustre HSM
2015/09/14-16:01:20.509073 lhsmtool_dmf[21731]: archive tier 2 default used
2015/09/14-16:01:20.509096 lhsmtool_dmf[21731]: archiving to Tier 2 (media/tape)
2015/09/14-16:01:20.509781 lhsmtool_dmf[21731]:
llapi_hsm_action_begin('/store/.lustre/fid/0x200010ccf:0x2c08:0x0')
2015/09/14-16:01:20.510224 lhsmtool_dmf[21731]: archive src: /store/.lustre/fid/0x200010ccf:0x2c08:0x0
2015/09/14-16:01:20.510248 lhsmtool_dmf[21731]: archive dst:
/archive/store2Backup/2c08/0000/0ccf/0001/0002/0000/0x200010ccf:0x2c08:0x0
2015/09/14-16:01:20.510257 lhsmtool_dmf[21731]: archive pth: CINES/cnu0003/toureille/lad15_demo/file1G
2015/09/14-16:01:20.510263 lhsmtool_dmf[21731]: archiving 'CINES/cnu0003/toureille/lad15_demo/file1G'
-> '/archive/store2Backup/2c08/0000/0ccf/0001/0002/0000/0x200010ccf:0x2c08:0x0' (Tier 2)
2015/09/14-16:01:20.537517 lhsmtool_dmf[21731]: dmu_archive_async
'/store/.lustre/fid/0x200010ccf:0x2c08:0x0' ->
'/archive/store2Backup/2c08/0000/0ccf/0001/0002/0000/0x200010ccf:0x2c08:0x0'
2015/09/14-16:01:20.537802 lhsmtool_dmf[21731]: dmu_archive_async dispatched dmureqid 17811
(...)
2015/09/14-16:03:43.956524 lhsmtool_dmf[21730]: archive 'CINES/cnu0003/toureille/lad15_demo/file1G'
(Tier 2) succeeded
2015/09/14-16:03:43.956554 lhsmtool_dmf[21730]: Action completed, notifying coordinator
cookie=0x5605081f, FID=[0x200010ccf:0x2c08:0x0], hp_flags=0 err=0
2015/09/14-16:03:43.957909 lhsmtool_dmf[21730]:
llapi_hsm_action_end('/store/.lustre/fid/0x200010ccf:0x2c08:0x0')
2015/09/14-16:03:43.957934 lhsmtool_dmf[21730]: SGITIME t2archive 18446744073709551615    143.450170
```

Check HSM state again :

```
narcisse-nfs2:/store/lad15_demo # lfs hsm_state ./file1G
./file1G: (0x00000009) exists archived, archive_id:1
```

The file is archived. Now let's try to gain some space

```
narcisse-nfs2:/store/lad15_demo # lfs hsm_state ./file1G
./file1G: (0x00000009) exists archived, archive_id:1
narcisse-nfs2:/store/toureille/lad15_demo # lfs hsm_release ./file1G
narcisse-nfs2:/store/toureille/lad15_demo # lfs hsm_state ./file1G
./file1G: (0x0000000d) released exists archived, archive_id:1
```

The hsm_release action is a synchronous process.

Restoring data

How to recover a deleted file on a Lustre-HSM cluster ?

With the /archive/shadow

```
narcisse-srv1:~ # ls /archive/shadow
ls: impossible d'accéder à /archive/shadow: Aucun fichier ou dossier de ce type
```

Since lhsmtool_dmf version 1.0.7, no more shadow directory...

Ok… so how can I restore files ?

Restoring data

Thanks to Robinhood and his «Deferred removal policy », when a file is delete, Robinhood saves the « Fid » and path into the SOFT_RM table.

```
mysql> SHOW COLUMNS FROM SOFT_RM;
+--------------+------------------+------+-----+---------+-------+
| Field        | Type             | Null | Key | Default | Extra |
+--------------+------------------+------+-----+---------+-------+
| fid          | varchar(64)      | NO   | PRI | NULL    |       |
| fullpath     | varchar(4095)    | YES  |     | NULL    |       |
| soft_rm_time | int(10) unsigned | YES  |     | NULL    |       |
| real_rm_time | int(10) unsigned | YES  | MUL | NULL    |       |
+--------------+------------------+------+-----+---------+-------+
4 rows in set (0.00 sec)
```

We can find our deleted file :

```
mysql> select fid,fullpath from SOFT_RM where fullpath like '%file1G%';
+-------------------------+-----------------------------------------------------+
| fid                     | fullpath                                            |
+-------------------------+-----------------------------------------------------+
| 0x200010ccf:0x2c08:0x0 | /store2/CINES/cnu0003/toureille/lad15_demo/file1G |
+-------------------------+-----------------------------------------------------+
1 row in set (0.08 sec)
```

Fortunately, we got the Fid 0x200010ccf:0x2c08:0x0

with which we can deduce the good path :

/archive/store2/2c08/*/0ccf/*/*/*/0x200010ccf:0x2c08:0x0

```
~ # echo  /archive/store2Backup/2c08/*/0ccf/*/*/*/0x200010ccf:0x2c08:0x0
/archive/store2Backup/2c08/0000/0ccf/0001/0002/0000/0x200010ccf:0x2c08:0x0

~ # cp   /archive/store2Backup/2c08/*/0ccf/*/*/*/0x200010ccf:0x2c08:0x0 \
/store/toureille/lad15_demo/file1G

~ # ls -l /store/CINES/lad15_demo/file1G
-rw-r----- 1 root root 1447447676 15 sept. 11:48 /store/toureille/lad15_demo/file1G
```

Conclusion

- HSM works fine
- 6 months late because of bugs
- Intel IML not supported by the integrator
- No more « easy » restore functionality

Our expectations

- « Easy » and efficient restore functionnality
- More tools able to manage tape libraries

Our questions

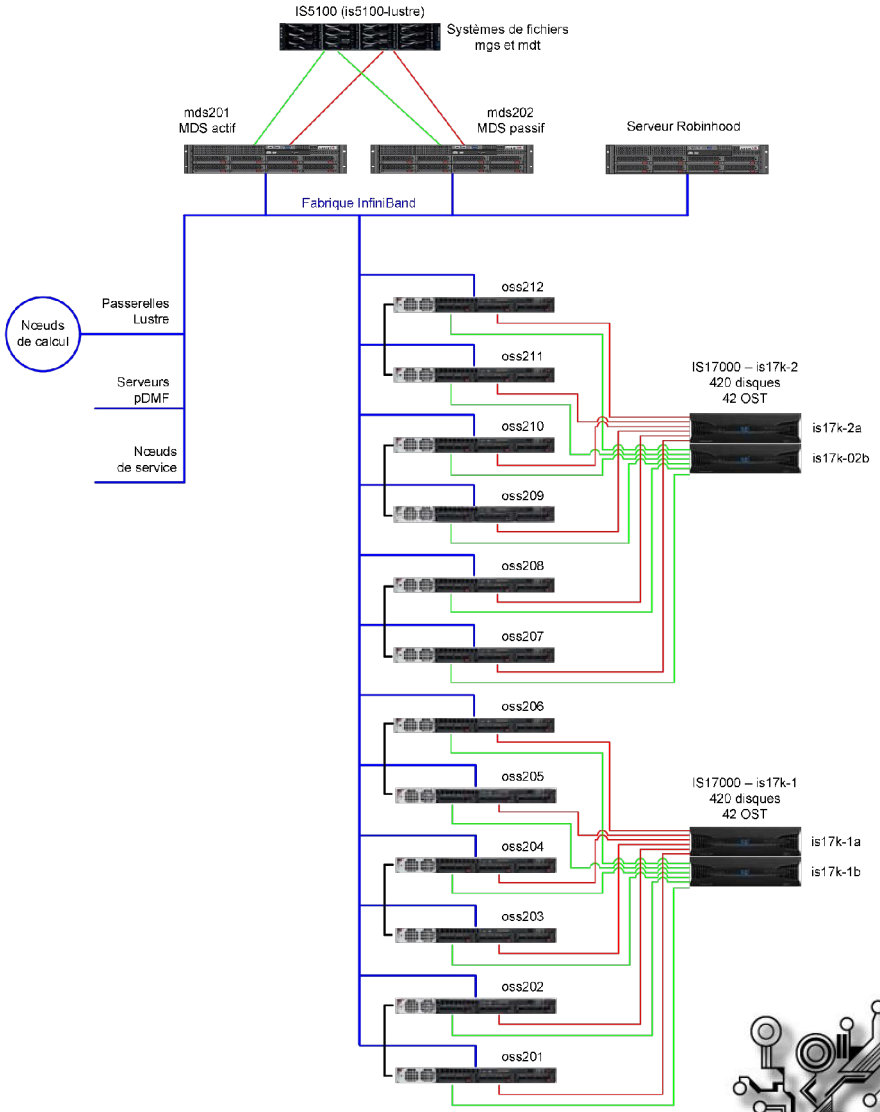
What about :

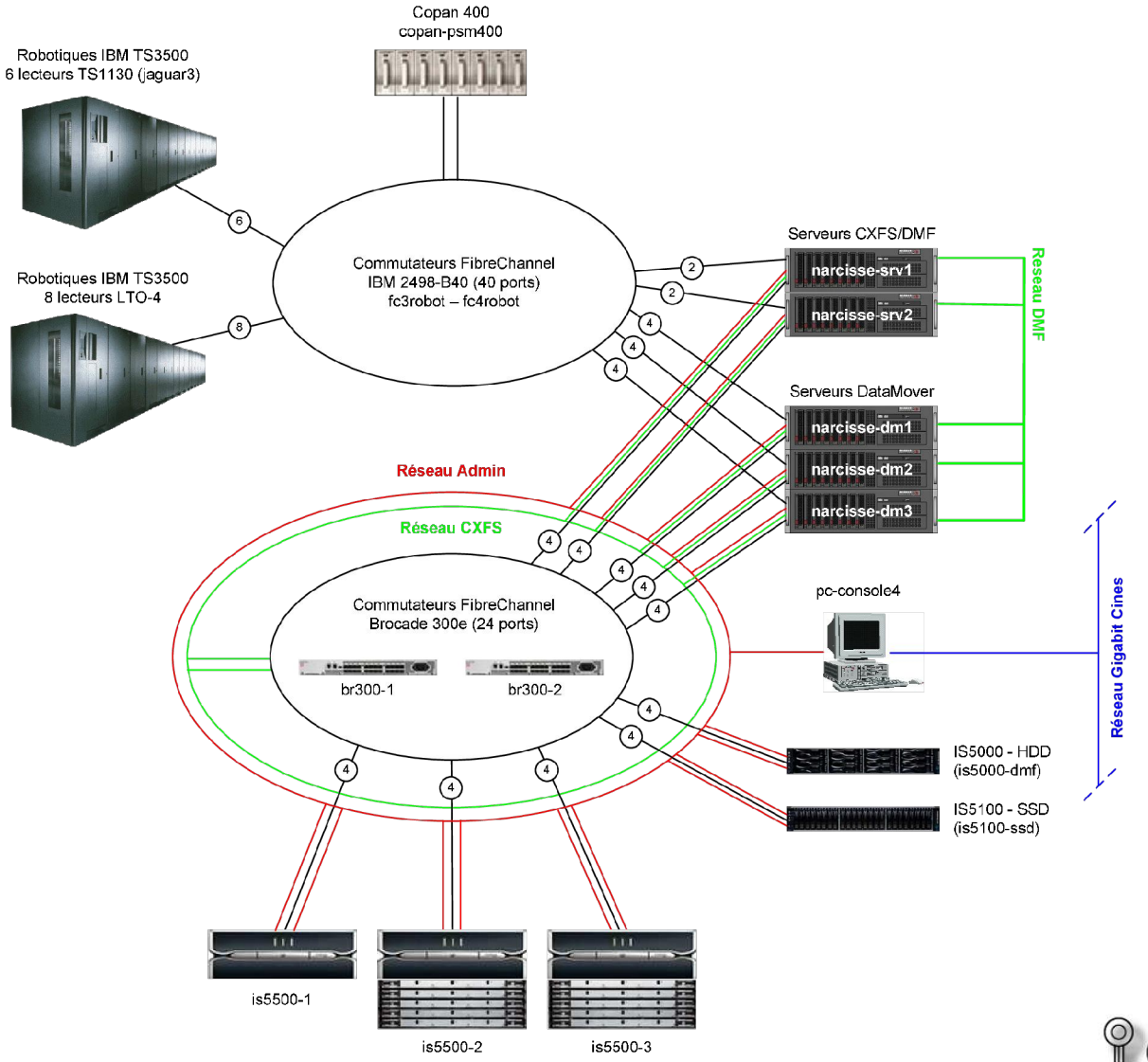
- future Lustre / RobinHood versions
- a better copytool integration
- the behavior when disk will be full

Thank you

Questions ?

Lustre HSM


IS5100 (is5100-lustre)
Systèmes de fichiers
mgs et mdt
mds201
MDS actif
mds202
MDS passif
Serveur Robinhood
Fabrique InfiniBand
Nœuds
de calcul
Passerelles
Lustre
Serveurs
pDMF
Nœuds
de service
oss212
oss211
oss210
oss209
oss208
oss207
oss206
oss205
oss204
oss203
oss202
oss201
IS17000 – is17k-2
420 disques
42 OST
is17k-2a
is17k-02b
IS17000 – is17k-1
420 disques
42 OST
is17k-1a
is17k-1b


Copan 400
copan-psm400
Robotiques IBM TS3500
6 lecteurs TS1130 (jaguar3)
Robotiques IBM TS3500
8 lecteurs LTO-4
Commutateurs FibreChannel
IBM 2498-B40 (40 ports)
fc3robot – fc4robot
Serveurs CXFS/DMF
narcisse-srv1
narcisse-srv2
Reseau DMF
Serveurs DataMover
narcisse-dm1
narcisse-dm2
narcisse-dm3
Réseau Admin
Réseau CXFS
Commutateurs FibreChannel
Brocade 300e (24 ports)
br300-1
br300-2
pc-console4
Réseau Gigabit Cines
IS5000 - HDD
(is5000-dmf)
IS5100 - SSD
(is5100-ssd)
is5500-1
is5500-2
is5500-3
/archive, /nfs05 et /nfs06

RobinHood
Migration policy

```
Migration_Parameters
{
[...]
    # interval for running migrations
    runtime_interval      = 8h ;

    # maximum number of migration requests per pass
    # (0: unlimited)
    max_migration_count   = 2000000 ;

    # maximum volume of migration requests per pass
    # (0: unlimited)
    max_migration_volume  = 8TB ;
[...]
}
```

RobinHood
Purge policy (Freeing space)

```
# Trigger purge on filesystem usage
Purge_Trigger
{
    trigger_on         = global_usage ;
    high_threshold_pct = 95% ;
    low_threshold_pct  = 90% ;
    check_interval     = 15min ;
    # raise an alert when the high threshold is reached
    alert_high         = TRUE ;
    # raise an alert if not enough data can be purged
    # to reach the low threshold
    alert_low          = TRUE ;
}
```

RobinHood
Remove policy (Hard delete)

```
# HSM remove policy (hard delete)
hsm_remove_policy
{
    # set this parameter to 'off' for disabling HSM object removal
    # hsm_remove = off;
    hsm_remove = enabled;
    # delay before impacting object removal in HSM
    deferred_remove_delay = 9d;
}
```