Deploying a LUSTRE filesystem for the HPC platform of the Research Area of the Bank of Italy.

Giuseppe Bruno
Bank of Italy

Lustre User Group, Austin Texas, U.S.A.
April 23-25

The views expressed are those of the author only and do not involve the responsibility of the Bank of Italy

OUTLINE

1. Motivation and focus of the paper
2. The Research Area and its computing needs
3. The employed hardware platform & LUSTRE architecture
4. Experimentation with three different file-systems
5. Storage Area Network
6. Main requirements for the filesystem
7. Lights and shadows: lack of knowledge vs software issues?
8. Concluding remarks

1 . Motivation and focus of the paper

The need to renew the computing platform for the Economic Research Area

Moving from a Shared memory to a distributed memory platform

Sharing a 10 TeraByte filesystem with 13 million files belonging to over 400 people logging on from different computing nodes

i. Can we provide a reliable and fast access to the filesystem?
ii. Can we provide a simple and transparent migration path for the users

2 . The Research Area and its computing needs

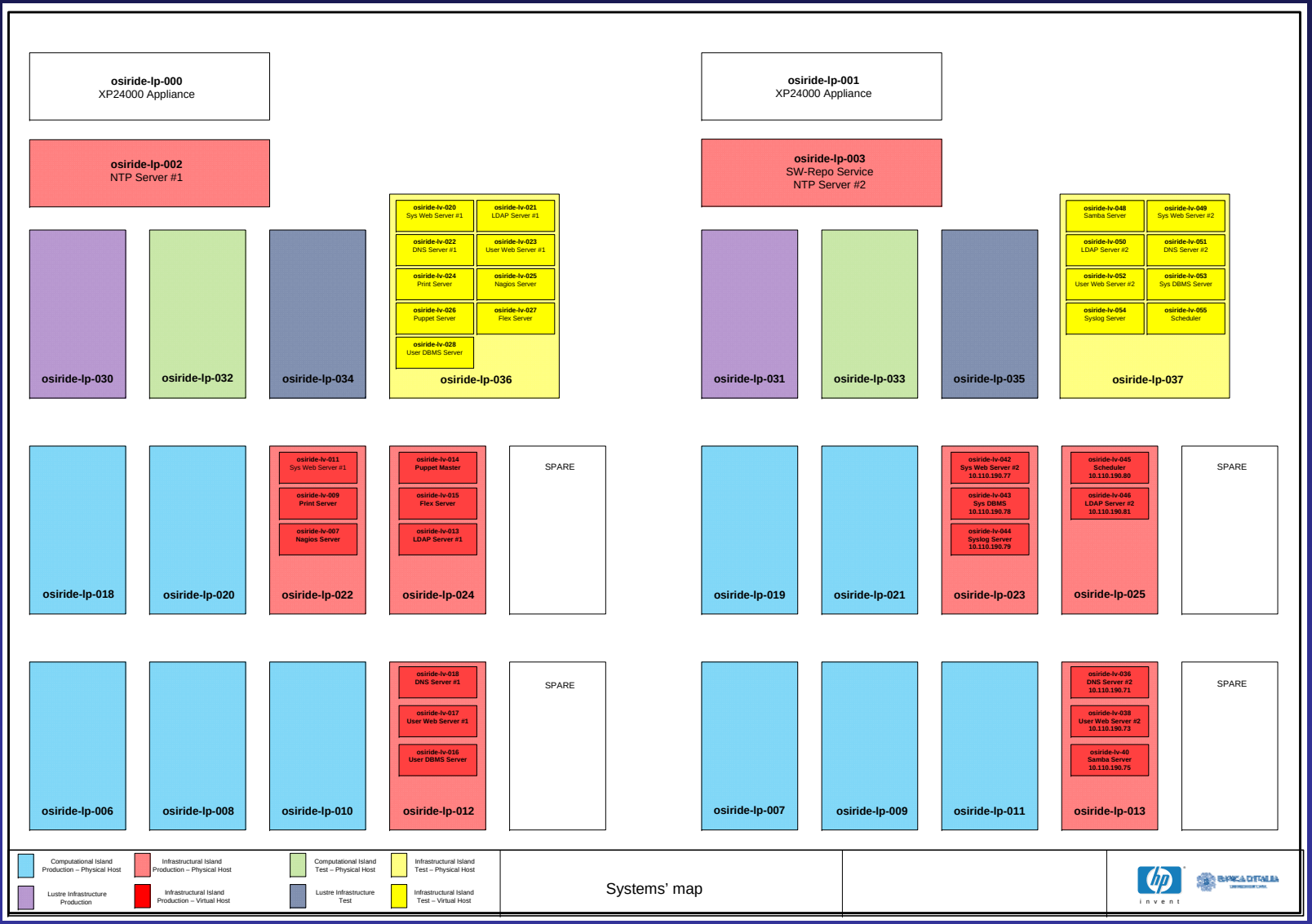
The B.I. Research Area is composed of more the 400 people. About another 100 are spread among 20 regional branches.

The researchers compile statistics, produce reports, run research projects and employ web applications.

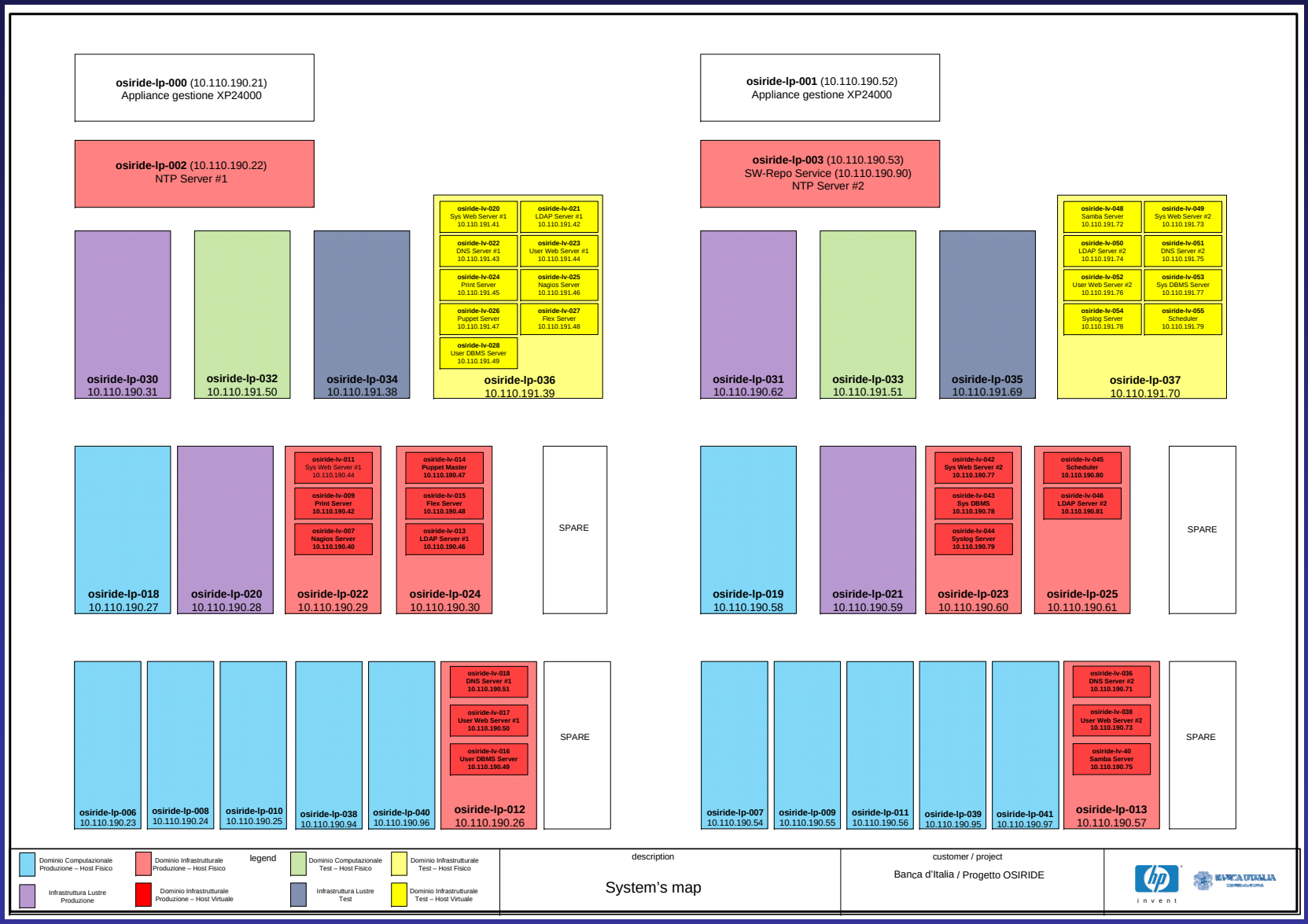
A relevant fraction of the computing tasks are interactive session with end-users packages:
MATLAB, SAS, Speakeasy, STATA, R …

Few users employ Fortran or C/C++ for parallel applications (openMP).

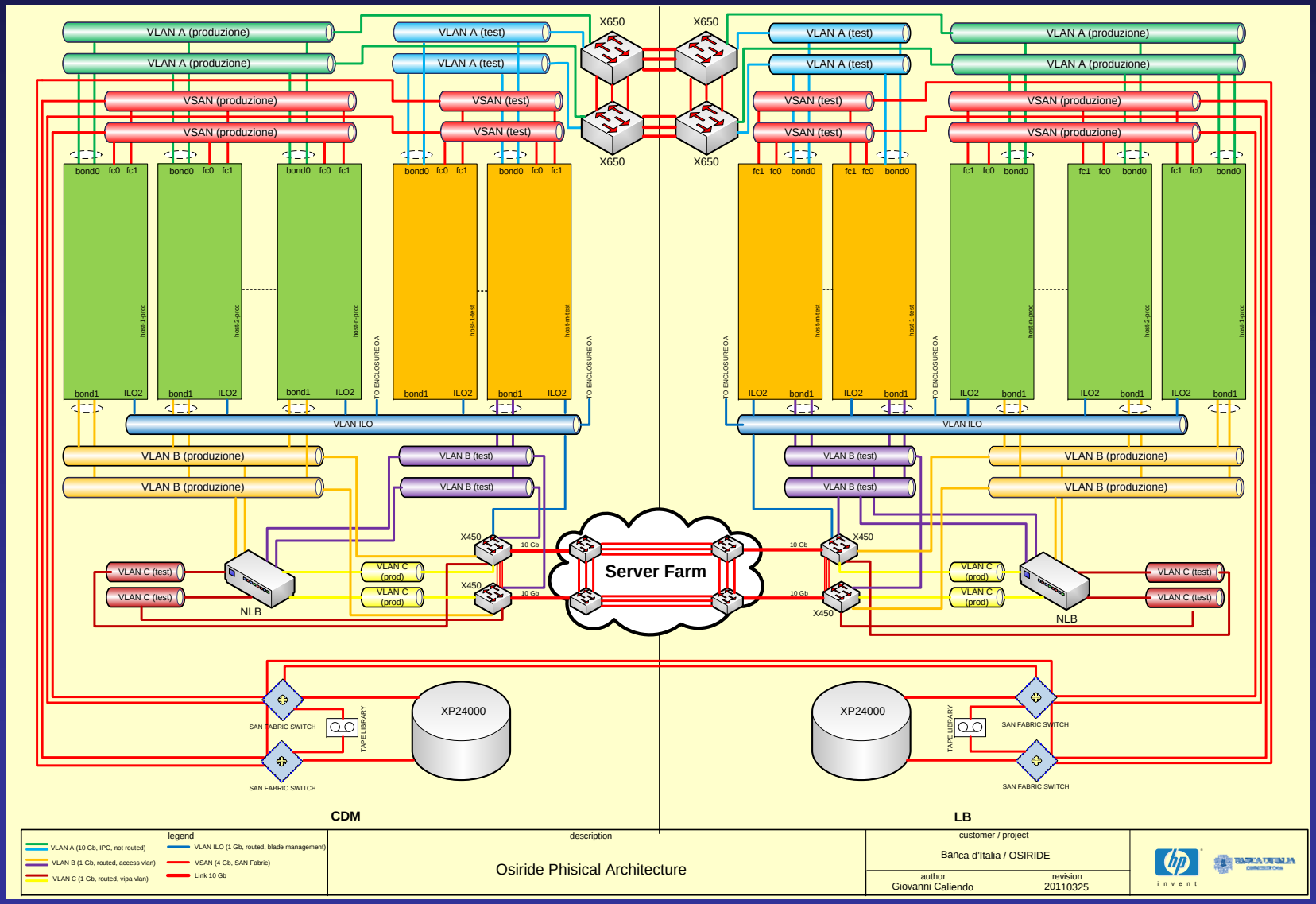
3. The starting hardware platform



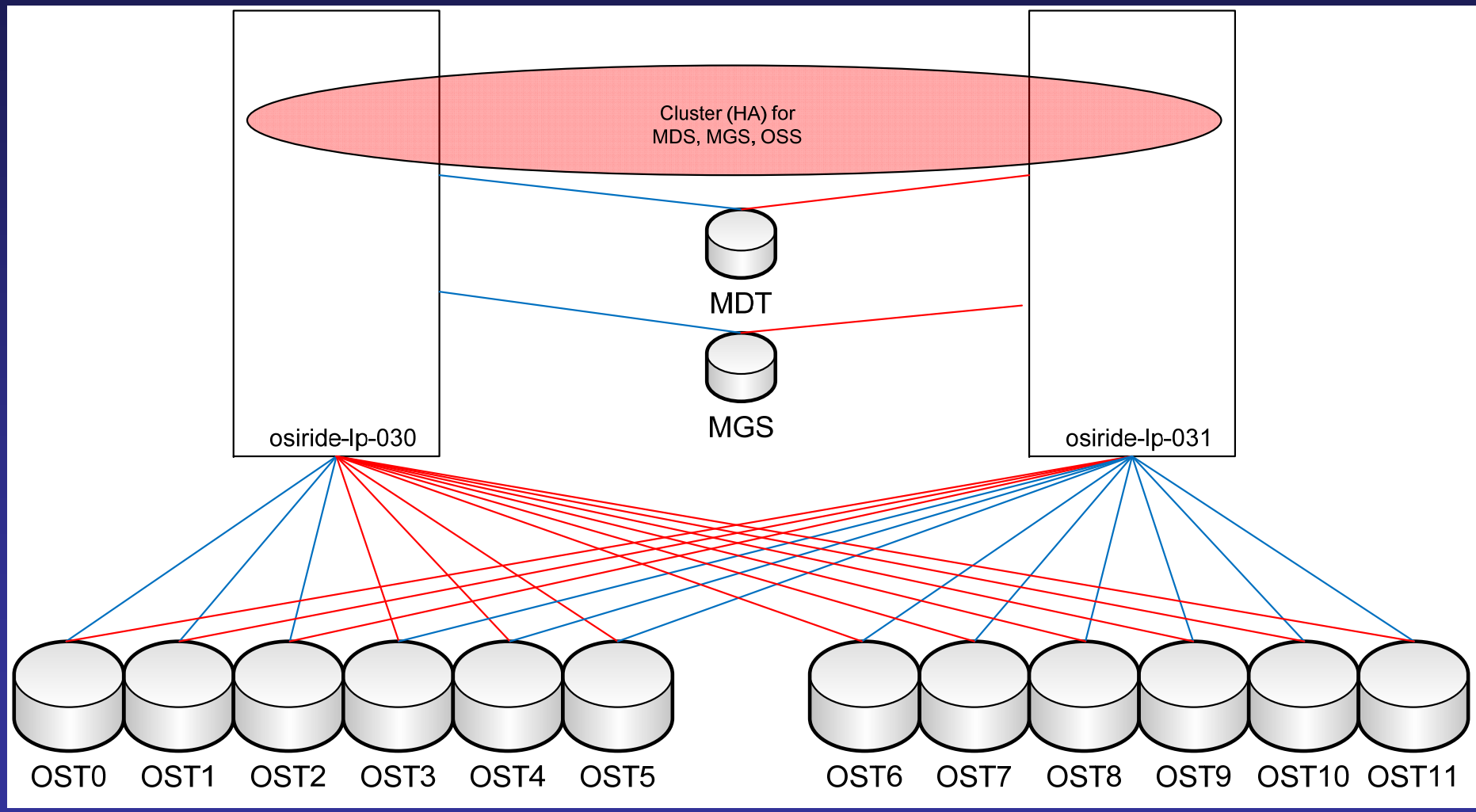
3. The upgraded hardware platform



3. The employed hardware platform



3. Initial Lustre server architecture

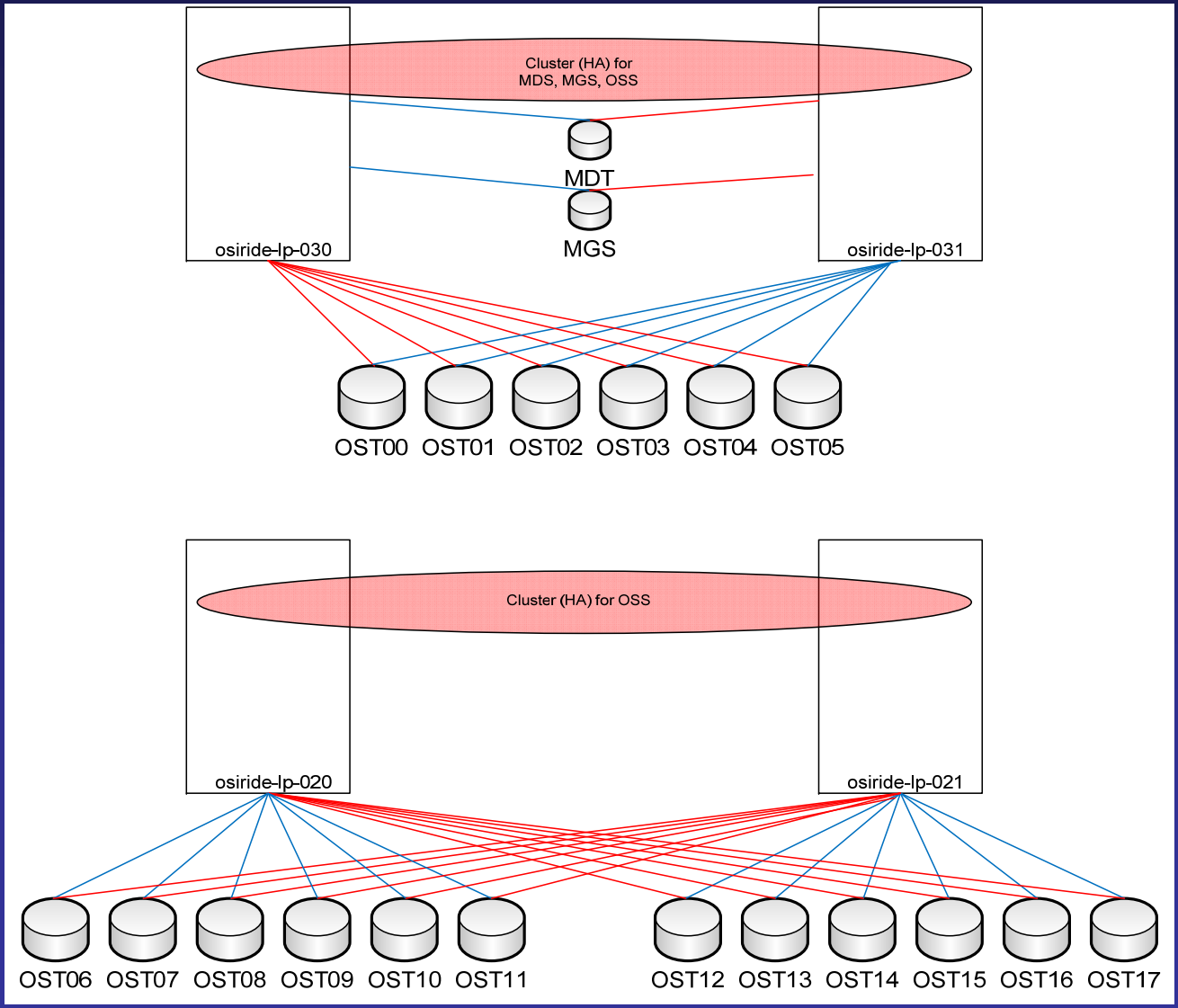


3. Initial Lustre server architecture

Member Name	ID	Status
osiride-lp-030-ipc.utenze.bankit.it	1	Online, Local, rgmanager
osiride-lp-031-ipc.utenze.bankit.it	2	Online, rgmanager
/dev/dm-45	0	Online, Quorum Disk

Service Name	Owner (Last)	State
service:ha_mdt	osiride-lp-030-ipc.utenze.bank	started
service:ha_mgs	osiride-lp-030-ipc.utenze.bank	started
service:ha_ost00	osiride-lp-030-ipc.utenze.bank	started
service:ha_ost01	osiride-lp-030-ipc.utenze.bank	started
service:ha_ost02	osiride-lp-030-ipc.utenze.bank	started
service:ha_ost03	osiride-lp-031-ipc.utenze.bank	started
service:ha_ost04	osiride-lp-031-ipc.utenze.bank	started
service:ha_ost05	osiride-lp-031-ipc.utenze.bank	started
service:ha_ost06	osiride-lp-031-ipc.utenze.bank	started
service:ha_ost07	osiride-lp-031-ipc.utenze.bank	started
service:ha_ost08	osiride-lp-031-ipc.utenze.bank	started
service:ha_ost09	osiride-lp-031-ipc.utenze.bank	started
service:ha_ost10	osiride-lp-031-ipc.utenze.bank	started
service:ha_ost11	osiride-lp-031-ipc.utenze.bank	started

3. Upgraded Lustre Server architecture



3. Upgraded Lustre Server architecture

Member Name	ID	Status
osiride-lp-030-ipc.utenze.bankit.it	1	Online, rgmanager
osiride-lp-031-ipc.utenze.bankit.it	2	Online, Local, rgmanager
/dev/dm-44	0	Online, Quorum Disk

Service Name	Owner (Last)	State
service:ha_mdt	osiride-lp-030-ipc.utenze.bank	started
service:ha_mgs	osiride-lp-030-ipc.utenze.bank	started
service:ha_ost00	osiride-lp-031-ipc.utenze.bank	started
service:ha_ost01	osiride-lp-031-ipc.utenze.bank	started
service:ha_ost02	osiride-lp-031-ipc.utenze.bank	started
service:ha_ost03	osiride-lp-031-ipc.utenze.bank	started
service:ha_ost04	osiride-lp-031-ipc.utenze.bank	started
service:ha_ost05	osiride-lp-031-ipc.utenze.bank	started

3. Upgraded Lustre Server architecture

Cluster Status for lustre-cluster2 @ Mon Apr 2 08:37:03 2012

Member Name	ID	Status
osiride-lp-020-ipc.utenze.bankit.it	1	Online, Local, rgmanager
osiride-lp-021-ipc.utenze.bankit.it	2	Online, rgmanager
/dev/dm-51	0	Online, Quorum Disk

Service Name	Owner (Last)	State
service:ha_ost06	osiride-lp-020-ipc.utenze.bank	started
service:ha_ost07	osiride-lp-020-ipc.utenze.bank	started
service:ha_ost08	osiride-lp-020-ipc.utenze.bank	started
service:ha_ost09	osiride-lp-020-ipc.utenze.bank	started
service:ha_ost10	osiride-lp-020-ipc.utenze.bank	started
service:ha_ost11	osiride-lp-020-ipc.utenze.bank	started
service:ha_ost12	osiride-lp-021-ipc.utenze.bank	started
service:ha_ost13	osiride-lp-021-ipc.utenze.bank	started
service:ha_ost14	osiride-lp-021-ipc.utenze.bank	started
service:ha_ost15	osiride-lp-021-ipc.utenze.bank	started
service:ha_ost16	osiride-lp-021-ipc.utenze.bank	started
service:ha_ost17	osiride-lp-021-ipc.utenze.bank	started

4. Experimentation with three different file-systems

Comparisons among GFSv2, OCFSv2 and LUSTRE v 1.8.3

We employed scripts for graphs management

File system	Tree creation	Tree update	Tree chgrp	Tree removal
GFS v2	150	146	2552	275
OCFS v2	361	172	37	181
Lustre v1.8	88	173	20	31
ext3 (local)	45	89	0.1	0.4

Times in seconds

5. Storage Area Network

1) The SAN is composed of a storage array including: 4 racks, each racks has 68 disks of 146 GByte for a raw total of 39 TBytes (4·68·146) (9 user TBytes for site);

2) Each set of 8 disks is in RAID5+HS configuration;
4 fabric switches in dual mirrored configuration;
In february 2012 we added 48 disks of 300 GByte for a total 14.4 TByte (7.2 user TBytes)

5. Storage Area Network

UUID	bytes	Used	Available	Use%	Mounted on
home-MDT0000_UUID	105.0G	3.4G	94.6G	3%	/home[MDT:0]
home-OST0000_UUID	708.7G	495.2G	177.5G	69%	/home[OST:0]
home-OST0001_UUID	708.7G	488.5G	184.2G	68%	/home[OST:1]
home-OST0002_UUID	708.7G	504.1G	168.6G	71%	/home[OST:2]
home-OST0003_UUID	708.7G	501.3G	171.4G	70%	/home[OST:3]
home-OST0004_UUID	708.7G	503.9G	168.8G	71%	/home[OST:4]
home-OST0005_UUID	708.7G	472.1G	200.6G	66%	/home[OST:5]
home-OST0006_UUID	708.7G	458.0G	214.7G	64%	/home[OST:6]
home-OST0007_UUID	708.7G	476.2G	196.5G	67%	/home[OST:7]
home-OST0008_UUID	708.7G	461.3G	211.4G	65%	/home[OST:8]
home-OST0009_UUID	708.7G	500.5G	172.2G	70%	/home[OST:9]
home-OST000a_UUID	708.7G	514.9G	157.8G	72%	/home[OST:10]
home-OST000b_UUID	708.7G	468.2G	204.5G	66%	/home[OST:11]
filesystem summary:	8.3T	5.7T	2.2T	68%	/home

5. Storage Area Network

UUID	Inodes	IUsed	IFree	IUse%	Mounted
home-MDT0000_UUID	39540771	12902083	26638688	32%	/home[MDT:0]
home-OST0000_UUID	47185920	1029415	46156505	2%	/home[OST:0]
home-OST0001_UUID	47185920	1062083	46123837	2%	/home[OST:1]
home-OST0002_UUID	47185920	967337	46218583	2%	/home[OST:2]
home-OST0003_UUID	47185920	1000791	46185129	2%	/home[OST:3]
home-OST0004_UUID	47185920	983731	46202189	2%	/home[OST:4]
home-OST0005_UUID	47185920	1057340	46128580	2%	/home[OST:5]
home-OST0006_UUID	47185920	1071393	46114527	2%	/home[OST:6]
home-OST0007_UUID	47185920	1035706	46150214	2%	/home[OST:7]
home-OST0008_UUID	47185920	1064001	46121919	2%	/home[OST:8]
home-OST0009_UUID	47185920	1004687	46181233	2%	/home[OST:9]
home-OST000a_UUID	47185920	948738	46237182	2%	/home[OST:10]
home-OST000b_UUID	47185920	1000120	46185800	2%	/home[OST:11]
filesystem summary:	39540771	12902083	26638688	32%	/home

5. Updated Storage Area Network

UUID	bytes	Used	Available	Use%	Mounted
home-MDT0000_UUID	105.0G	3.4G	94.6G	3%	/home[MDT:0]
home-OST0000_UUID	708.7G	487.4G	185.3G	72%	/home[OST:0]
home-OST0001_UUID	708.7G	472.7G	200.0G	70%	/home[OST:1]
home-OST0002_UUID	708.7G	421.4G	251.4G	63%	/home[OST:2]
home-OST0003_UUID	708.7G	473.2G	199.5G	70%	/home[OST:3]
home-OST0004_UUID	708.7G	482.2G	190.5G	72%	/home[OST:4]
home-OST0005_UUID	708.7G	459.6G	213.1G	68%	/home[OST:5]
home-OST0006_UUID	708.7G	466.2G	206.5G	69%	/home[OST:6]
home-OST0007_UUID	708.7G	457.8G	214.9G	68%	/home[OST:7]
home-OST0008_UUID	708.7G	470.2G	202.5G	70%	/home[OST:8]
home-OST0009_UUID	708.7G	454.4G	218.3G	68%	/home[OST:9]
home-OST000a_UUID	708.7G	436.5G	236.2G	65%	/home[OST:10]
home-OST000b_UUID	708.7G	460.4G	212.3G	68%	/home[OST:11]
home-OST000c_UUID	717.3G	21.6G	659.7G	3%	/home[OST:12]
home-OST000d_UUID	717.3G	11.6G	669.7G	2%	/home[OST:13]
home-OST000e_UUID	717.3G	15.1G	666.2G	2%	/home[OST:14]
home-OST000f_UUID	717.3G	15.7G	665.6G	2%	/home[OST:15]
home-OST0010_UUID	717.3G	13.5G	667.8G	2%	/home[OST:16]
home-OST0011_UUID	717.3G	16.0G	665.3G	2%	/home[OST:17]
filesystem summary:	12.5T	5.5T	6.4T	46%	/home

5. Updated Storage Area Network

UUID	Inodes	IUsed	IFree	IUse%	Mounted
home-MDT0000_UUID	73400320	13166391	60233929	18%	/home[MDT:0]
home-OST0000_UUID	47185920	1044005	46141915	2%	/home[OST:0]
home-OST0001_UUID	47185920	1076198	46109722	2%	/home[OST:1]
home-OST0002_UUID	47185920	986221	46199699	2%	/home[OST:2]
home-OST0003_UUID	47185920	1010920	46175000	2%	/home[OST:3]
home-OST0004_UUID	47185920	994924	46190996	2%	/home[OST:4]
home-OST0005_UUID	47185920	1070133	46115787	2%	/home[OST:5]
home-OST0006_UUID	47185920	1088727	46097193	2%	/home[OST:6]
home-OST0007_UUID	47185920	1050328	46135592	2%	/home[OST:7]
home-OST0008_UUID	47185920	1080625	46105295	2%	/home[OST:8]
home-OST0009_UUID	47185920	1022622	46163298	2%	/home[OST:9]
home-OST000a_UUID	47185920	958733	46227187	2%	/home[OST:10]
home-OST000b_UUID	47185920	1017007	46168913	2%	/home[OST:11]
home-OST000c_UUID	11151360	13727	11137633	0%	/home[OST:12]
home-OST000d_UUID	11151360	13644	11137716	0%	/home[OST:13]
home-OST000e_UUID	11151360	13744	11137616	0%	/home[OST:14]
home-OST000f_UUID	11151360	13605	11137755	0%	/home[OST:15]
home-OST0010_UUID	11151360	13514	11137846	0%	/home[OST:16]
home-OST0011_UUID	11151360	13677	11137683	0%	/home[OST:17]
filesystem summary:	73400320	13166391	60233929	18%	/home

6. Main requirements for the filesystem

About 13 million files. The size distribution is highly left skewed with a median of 4 Kb and average of 500 Kb.

Stripe parameters (small files)

lmm_stripe_count: 1

lmm_stripe_size: 1 MByte

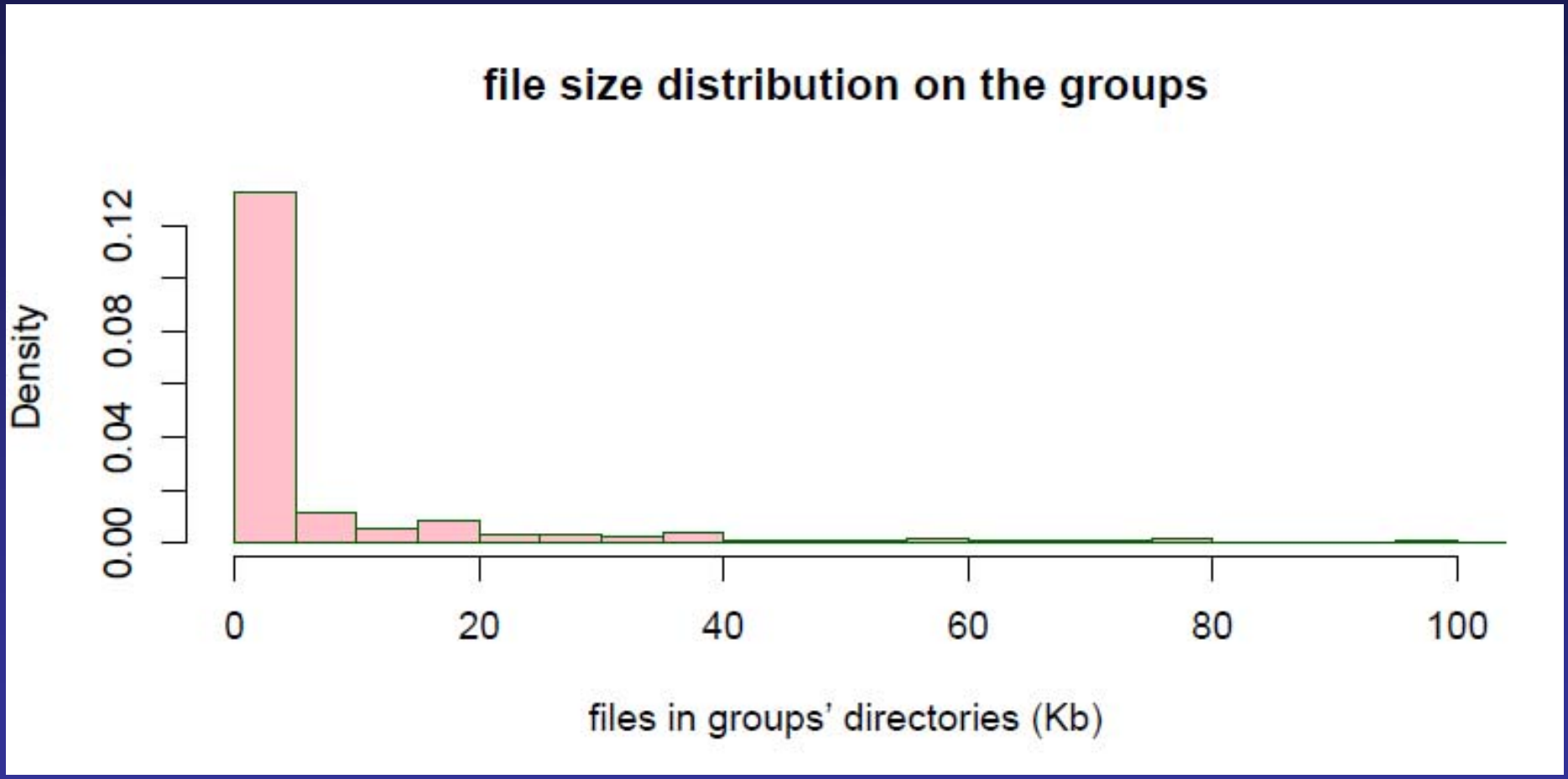
Free Space & Stripe allocation parameters

qos_prio_free : 89%

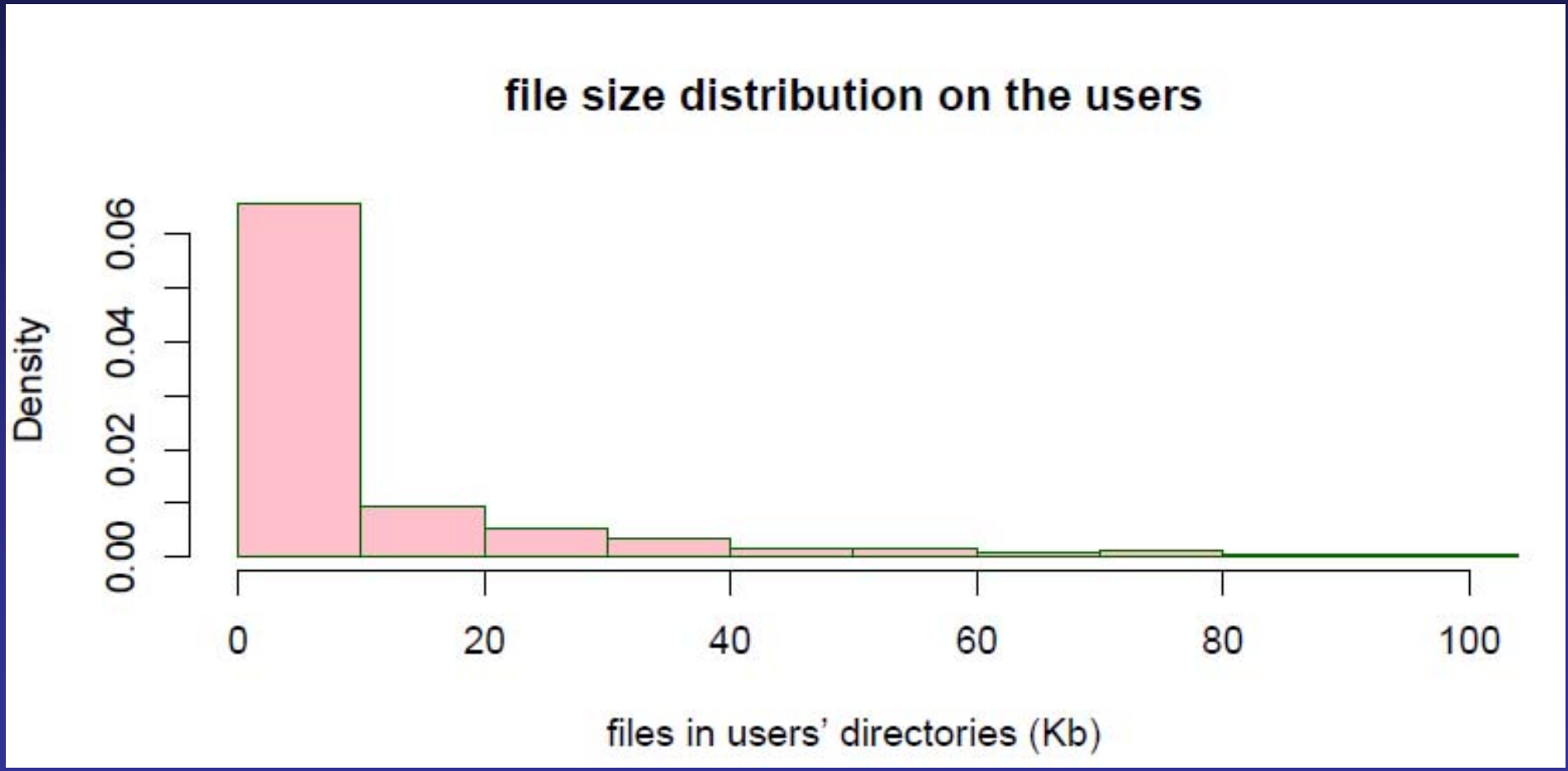
qos_threshold_rr : 16%

These parameters should speed up the OSTs balancing

6. Main requirements for the filesystem



6. Main requirements for the filesystem



6. Main requirements for the filesystem

```
[osiride-lp-031 ~]$ /usr/sbin/lctl get_param ost.OSS.ost.threads_*
ost.OSS.ost.threads_max=512
ost.OSS.ost.threads_min=128
ost.OSS.ost.threads_started=512
[osiride-lp-020 ~]$ /usr/sbin/lctl get_param ost.OSS.ost.threads_*
ost.OSS.ost.threads_max=512
ost.OSS.ost.threads_min=128
ost.OSS.ost.threads_started=128
[osiride-lp-021 ~]$ /usr/sbin/lctl get_param ost.OSS.ost.threads_*
ost.OSS.ost.threads_max=512
ost.OSS.ost.threads_min=128
ost.OSS.ost.threads_started=512
[osiride-lp-030 lustre]$ /usr/sbin/lctl get_param mdt.MDS.mds.threads_*
mdt.MDS.mds.threads_max=128
mdt.MDS.mds.threads_min=32
mdt.MDS.mds.threads_started=32
```

7. Lights and shadows: lack of knowledge vs software issues?

1.8.3 v Roll-out in february 2011

chmod misbehaviour in directory with sgid on

SAMBA opportunistic locks are sometimes not honoured incompatibility between LUSTRE and SAMBA locks

Migration to 1.8.5 in June 2011

frequent crash of one of the LUSTRE server hosting MGS/MDT.

Difficulty in managing quotas distributed among OSTs

7. Lights and shadows: lack of knowledge vs software issues?

SAMBA tries to use LOCK_MAND instead of LOCK_EX

LUSTRE doesn’t recognizes LOCK_MAND (x32)

osiride-lp-010 kernel: LustreError: 14384:0: (file.c:3227: ll_file_flock()) unknown fcntl lock type: 32

Lustre Manual: 9.1.4 **Known Issues with Quotas**

Using quotas in Lustre can be complex and there are several known issues.

7. Lights and shadows: lack of knowledge vs software issues?

Modification of

lquota.*.quota_switch_qs = 0

(dynamically changing qunit size is disabled)

Addition of 6 new OSTs

Migration to 1.8.7wc in march 2012

From January 2012 the platform seems stabilized.

8. Concluding Remarks

Deploying a Lustre filesystem with many small files is quite challenging;

Lustre filesystem has shown very good performance features,

good performances didn’t match with resilience and stability,

suboptimal configuration? Unfit design with our filesystem?

SAMBA server doesn’t seem to fit smoothly on LUSTRE filesystem;

Managing user/group Quotas might probably be improved by offering quotas tighter to user needs.

Thank you for your attention.

Giuseppe Bruno
Bank of Italy
Research and International Relations
Head of I.T. Support Unit
giuseppe.bruno@bancaditalia.it