Lustre Experience at CEA/DIF

J-Ch Lafoucrière

jc.lafoucriere@cea.fr

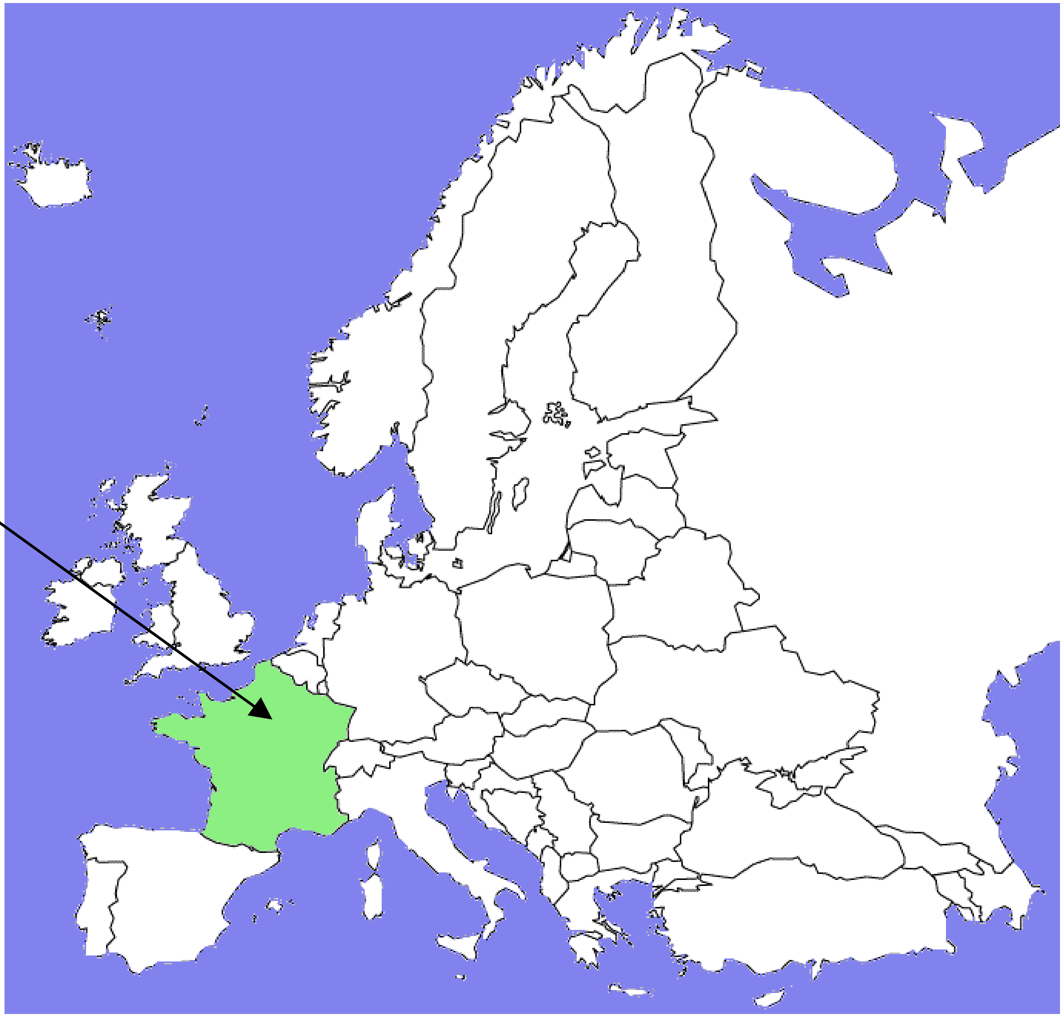
- CEA/DIF Computing Center
- Lustre File System
- CEA/DIF Lustre Configurations and Performances

CEA/DIF Computing Center

CEA Computing Center

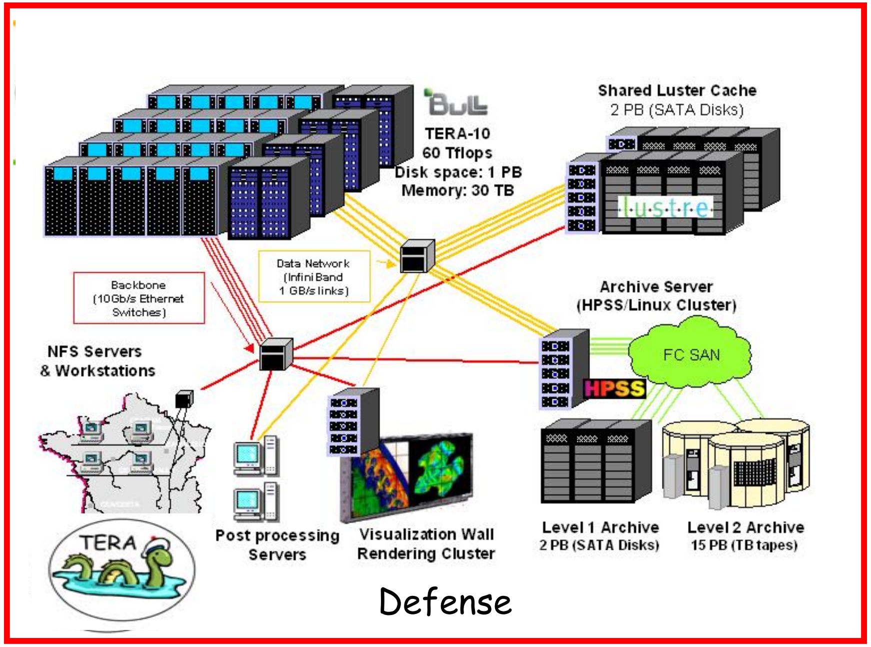
CEA DIF

(South of Paris)



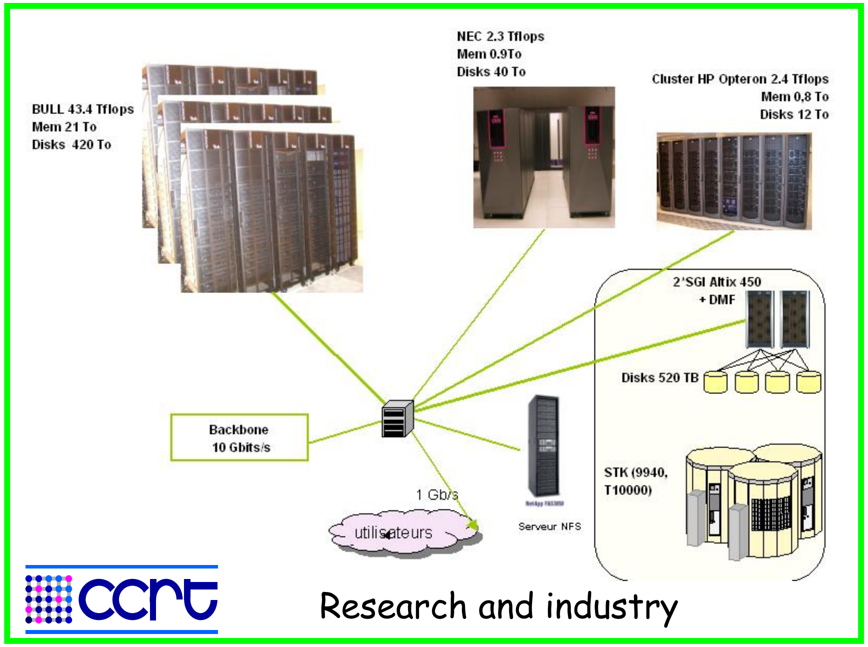
Current supercomputers at CEA/DIF

TERA-10



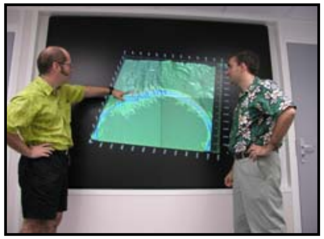
Defense

CCRT-B

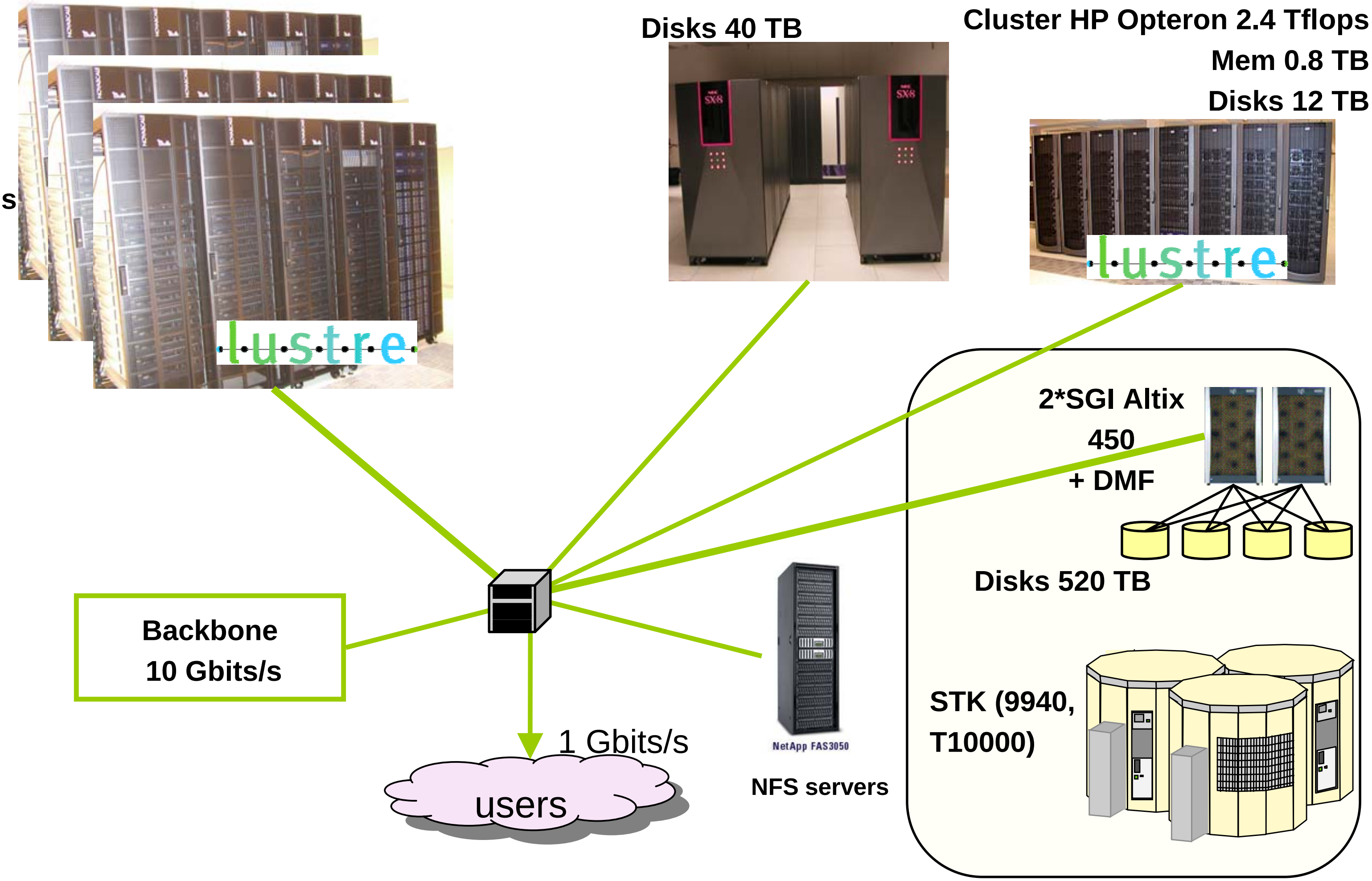


Research and industry

New
Technologies



CCRT-B Computing Center Architecture



BULL 43.4 Tflops
Mem 21 TB
Disks 420 TB

NEC 2.3 Tflops
Mem 0.9TB
Disks 40 TB

Cluster HP Opteron 2.4 Tflops
Mem 0.8 TB
Disks 12 TB

2*SGI Altix
450
+ DMF

Disks 520 TB

STK (9940,
T10000)

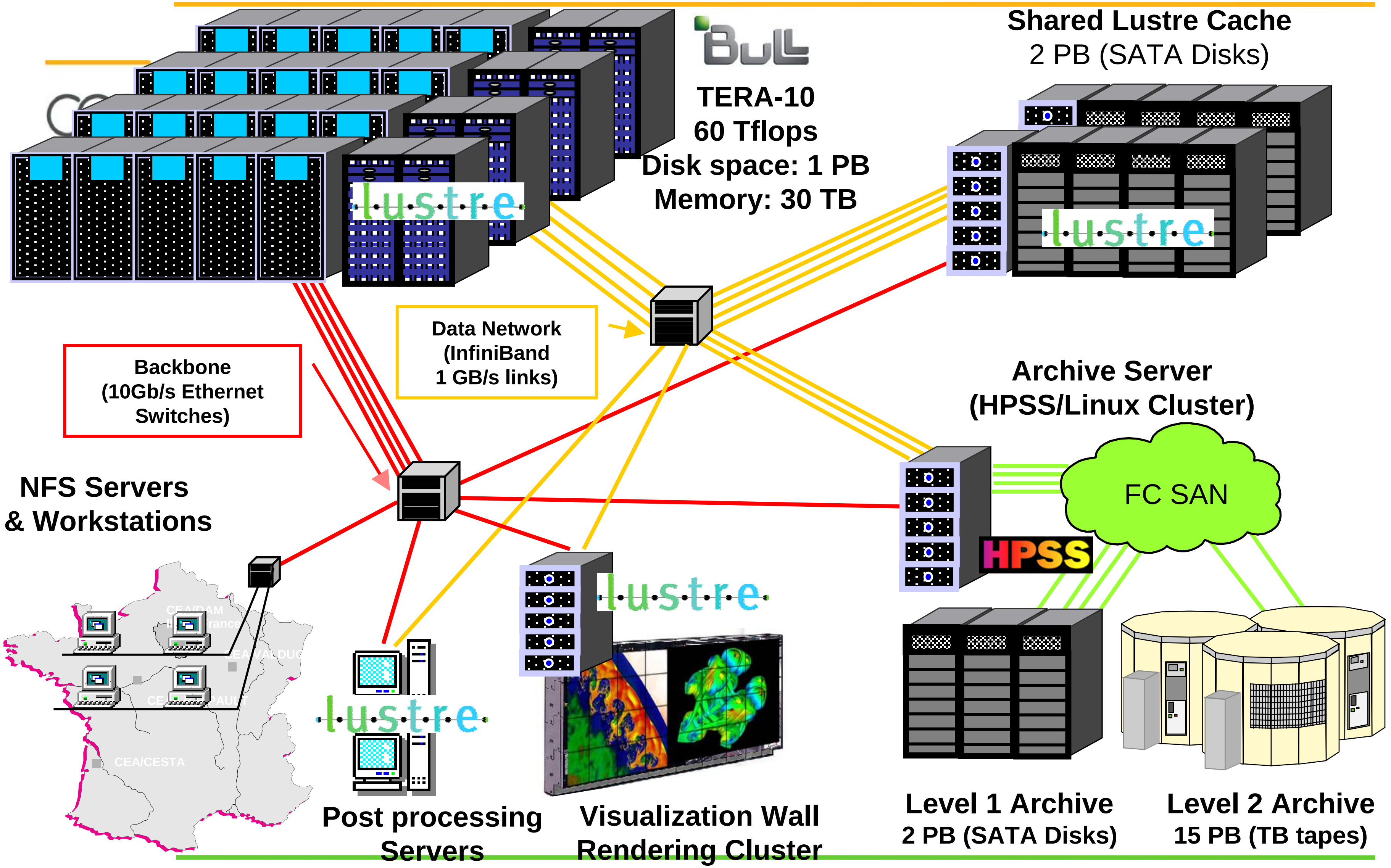
Backbone
10 Gbits/s

1 Gbits/s

users

NFS servers

TERA Computing Center Architecture



TERA-10
60 Tflops
Disk space: 1 PB
Memory: 30 TB

Shared Lustre Cache
2 PB (SATA Disks)

Data Network
(InfiniBand
1 GB/s links)

Backbone
(10Gb/s Ethernet
Switches)

Archive Server
(HPSS/Linux Cluster)

FC SAN

NFS Servers
& Workstations

Post processing
Servers

Visualization Wall
Rendering Cluster

Level 1 Archive
2 PB (SATA Disks)

Level 2 Archive
15 PB (TB tapes)

Lustre File System

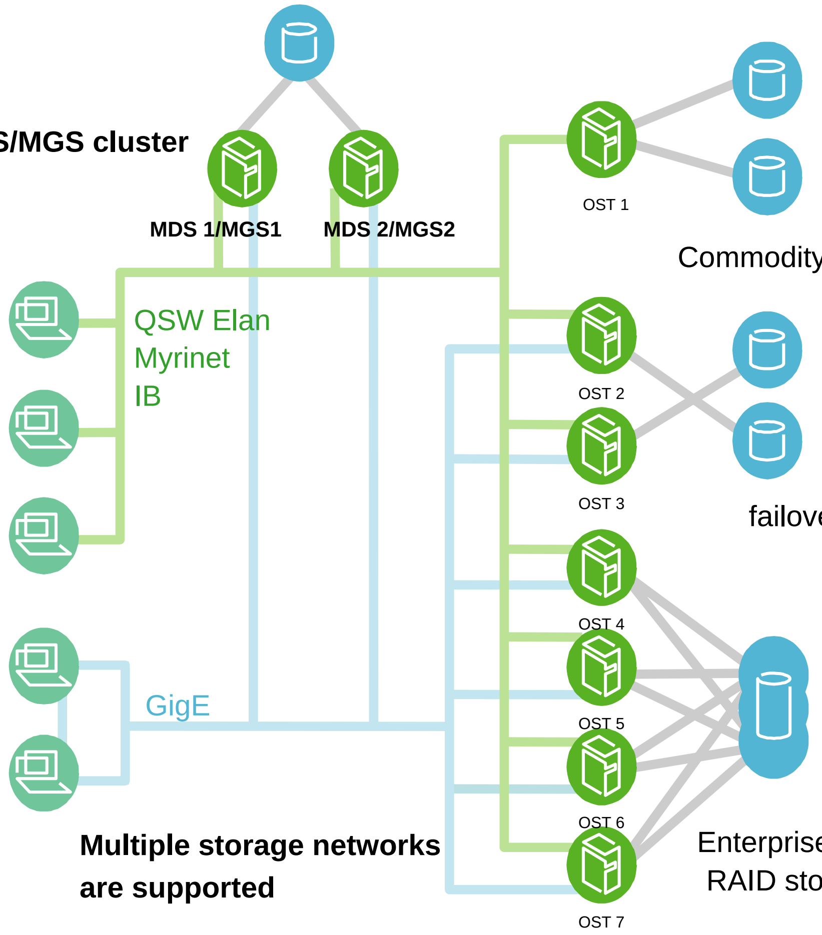
What’s Lustre ?

- **A high performance filesystem**
  - A new storage architecture (storage object)
  - Designed for performances
    - X0 000 nodes, Peta bytes of storage, large directories, …
    - 90 % hardware efficiency

- **Open Source Project**
  - Available as tarball and rpm from CFS (RH, Suse)
    - All tools available to make site specific rpm
    - 2.4 and 2.6 Linux kernels support
  - Available through vendors integration (HP, LNXI, Cray, Bull, IBM, SUN, …)

- **Managed by CFS as a product, not as a best effort project**

Lustre Cluster



MDS/MGS cluster

MDS 1/MGS1

MDS 2/MGS2

OST 1

Commodity SAN or disks

QSW Elan
Myrinet
IB

OST 2

OST 3

failover

Pools of
Lustre OSS
(1000’s)

OST 4

GigE

OST 5

OST 6

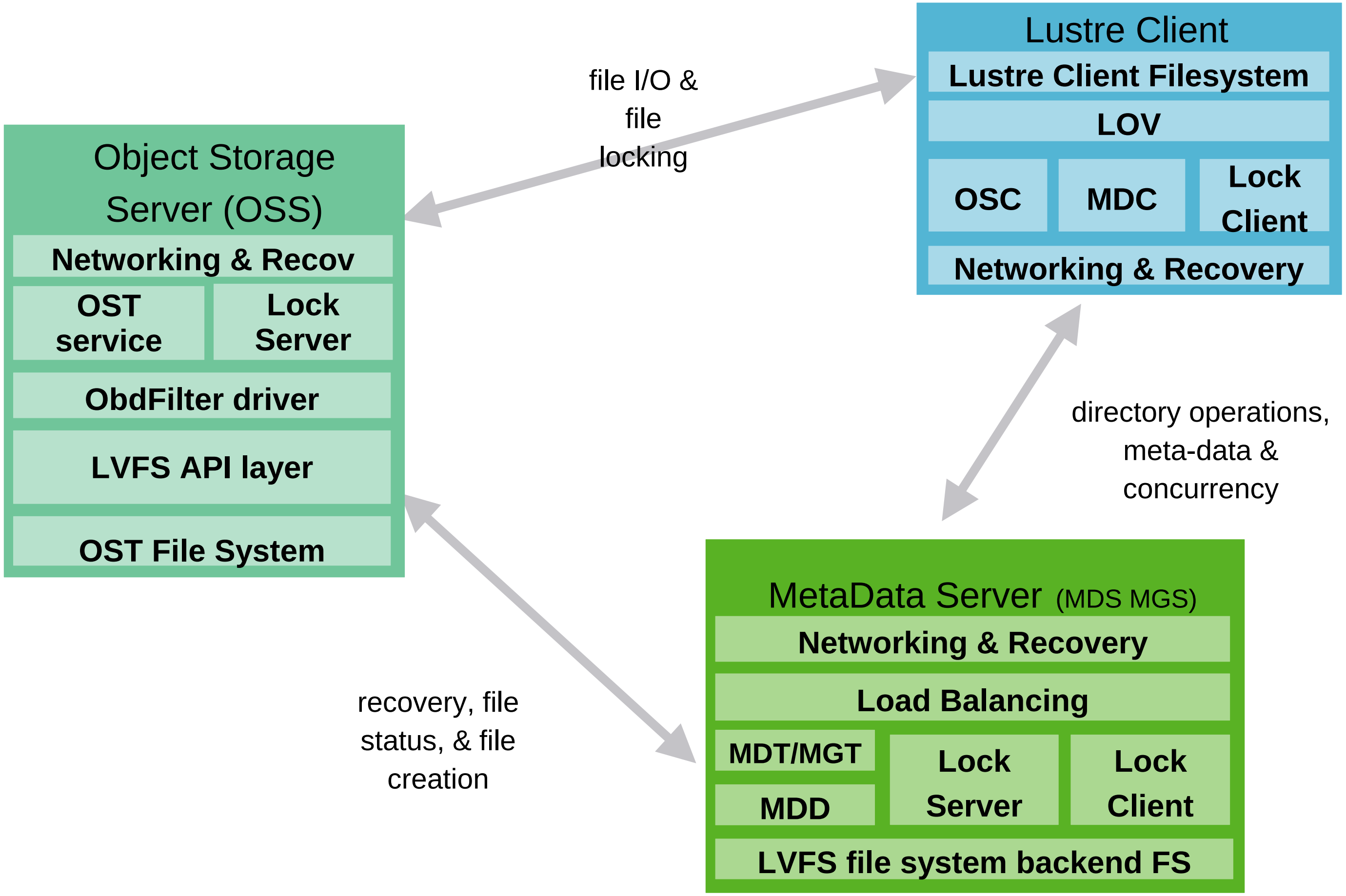
OST 7

Clients
Up to X0,000’s

Multiple storage networks
are supported

Enterprise-class
RAID storage

Lustre Components



Lustre Client

Lustre Client Filesystem

LOV

OSC

MDC

Lock Client

Networking & Recovery

file I/O & file locking

Object Storage Server (OSS)

Networking & Recov

OST service

Lock Server

ObdFilter driver

LVFS API layer

OST File System

directory operations, meta-data & concurrency

recovery, file status, & file creation

MetaData Server (MDS MGS)

Networking & Recovery

Load Balancing

MDT/MGT

MDD

Lock Server

Lock Client

LVFS file system backend FS

Lustre Design Rules

- **All software uses stackable modules**
  - Storage devices are accessed through a local filesystem ldiskfs (an ext3 based FS, very close to ext4) and others in the future (ZFS ?)
  - Network layer (LNET) is a message passing library
    - Hardware independence
    - Transactional RPC or Bulk transfers, use of callbacks
  - Networks can be heterogeneous with LNET routers
- **IO performances**
  - Large I/O sizes on networks and storage devices
  - Highly parallel
  - Large client cache
- **Metadata performances**
  - Fine and dynamic lock granularity
- **Robust design**
  - High Availability + journalization

Today Status

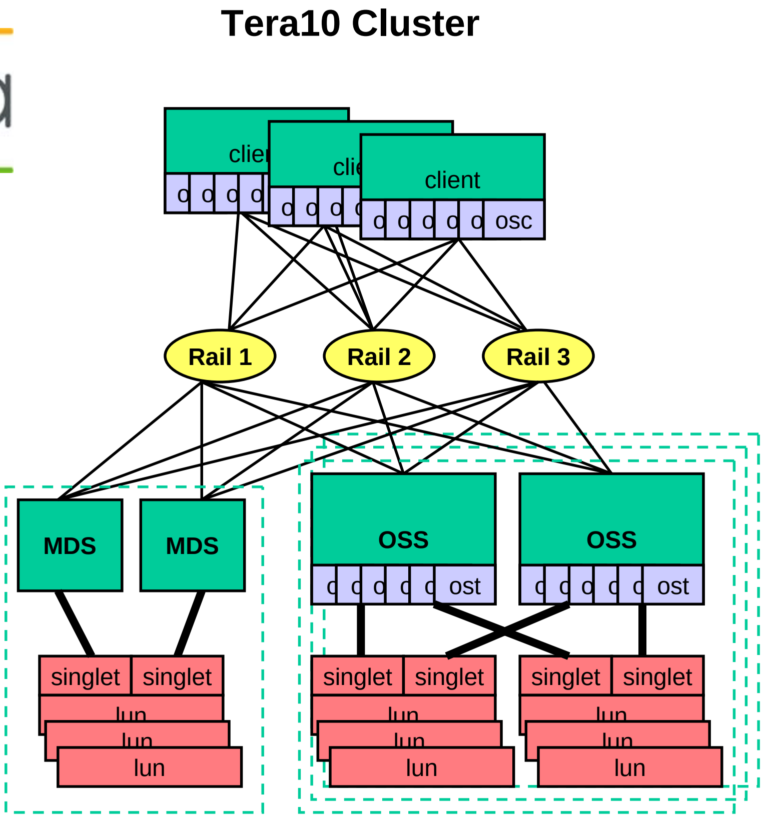
- **Last release is 1.4.10**
  - Scalable product
  - ACL, Extended Attributes, Quotas
  - Fault tolerant
- **Next major release 1.6.0 (these days)**
  - New configuration tools: Only 2 commands, mkfs and mount
  - Online extension
  - Patchless client support
  - Device multimount protection
  - Large OBD (8 TB)

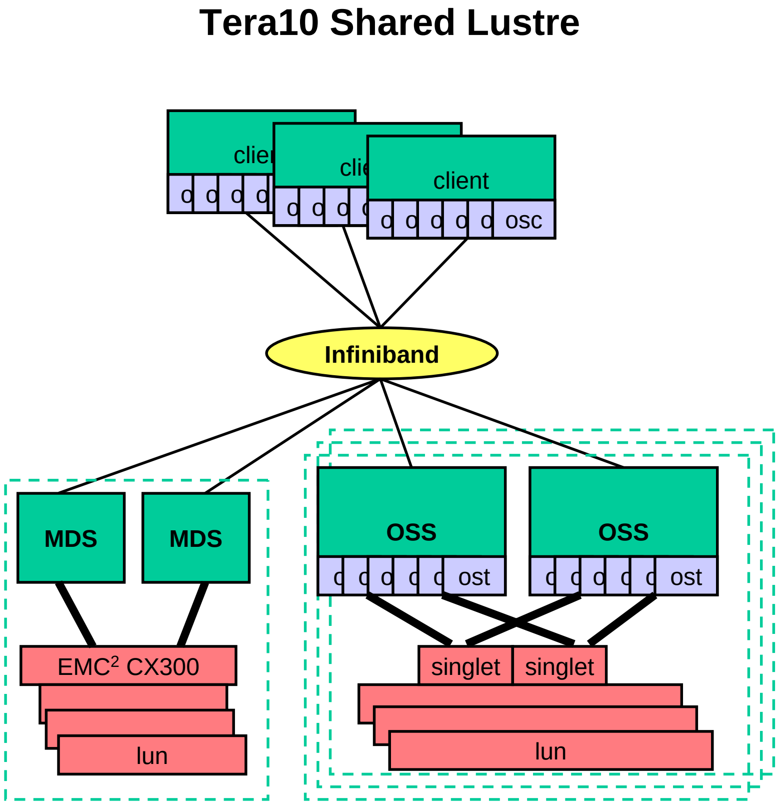
Evolutions

- **End 2007, 2008 (1.6.x, 1.8)**
  - Storage pools
  - Lustre OST RAID
  - Kerberos support
- **Ports to non Linux platforms are planed (may be non OpenSource)**
  - Servers port on Solaris will be in userspace
- **Servers may move to userspace on Linux**

CEA/DIF Lustre Configurations and Performances

Two Lustre architectures

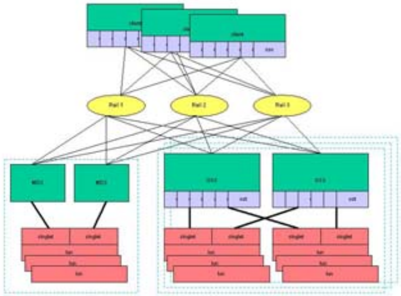




- Quadrics Elan4 interconnect, 3 rails each 900 MB/s
- 565 clients
- IO Cell = 2 OSS + 2 DDN couplet

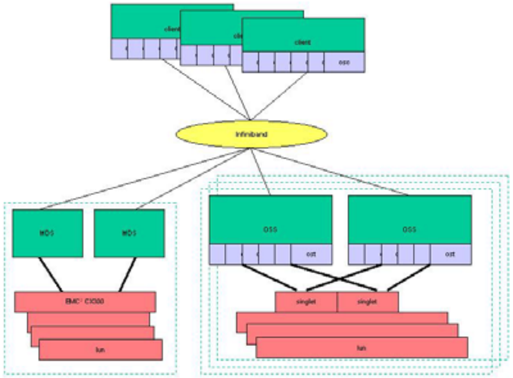
- Infiniband SDR interconnect
- ~60 clients
- IO Cell = 2 OSS + 1 DDN couplet

Lustre Usage (I)



- **HPC Cluster File System**
  - Lustre 1.4.7 or Lustre 1.4.8, with vendor support
  - Quadrics (Elan4) or InfiniBand (DDR) networks
    - Native network protocols (qsnet/OpenIB gen2 LND)
    - Network is dedicated to a cluster
  - OSS/MDS are 16 or 8 Itanium cores servers
  - Dual attached DDN 9550 with fibre channel disks (8D+1P+1S, writethrough), 16 LUN of 1TB
  - 4.3 GB/s per IO Cell (2 DDN couplet)
  - Mounted by one cluster (few FS per cluster)
  - Performance oriented
    - 100 GB/s on checkpoint/restart like benchmark
    - Single client performance: 2.2 GB/s W, 1.4 GB/s R
  - TERA-10 day production is 30 TB

Lustre Usage (II)

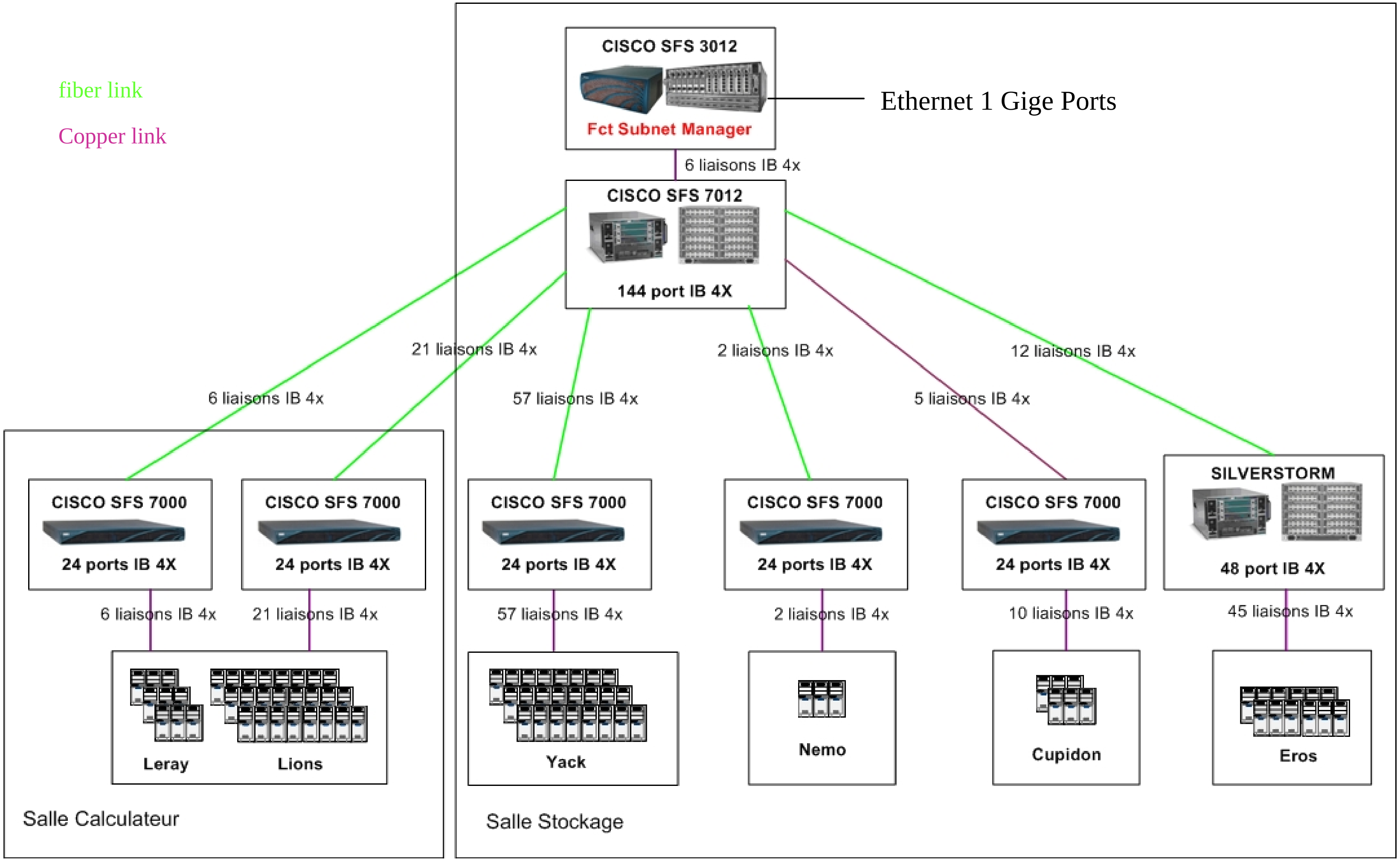


- **Shared File system**
  - Lustre 1.6 Beta 7
  - InfiniBand network (SDR)
    - Native network protocol (OpenIB gen2 LND)
    - Network is shared by clusters
  - OSS/MDS are 4 Xeon core servers
  - Dual attached DDN 9550 with SATA disks (8D+2P+shared S), writethrough), 48 LUN of 8 TB
  - 1.5 GB/s per IO Cell (1 DDN couplet), limit is the 2 IB links
  - 1 FS shared by multiple clusters
  - Capacity oriented:
    - multi peta bytes FS (target is 2+ PB)
    - Single client performance: 472 MB/s W, 371 MB/s R
      - Limit is the client machine (only 2 cores)

Shared InfiniBand Network Topology

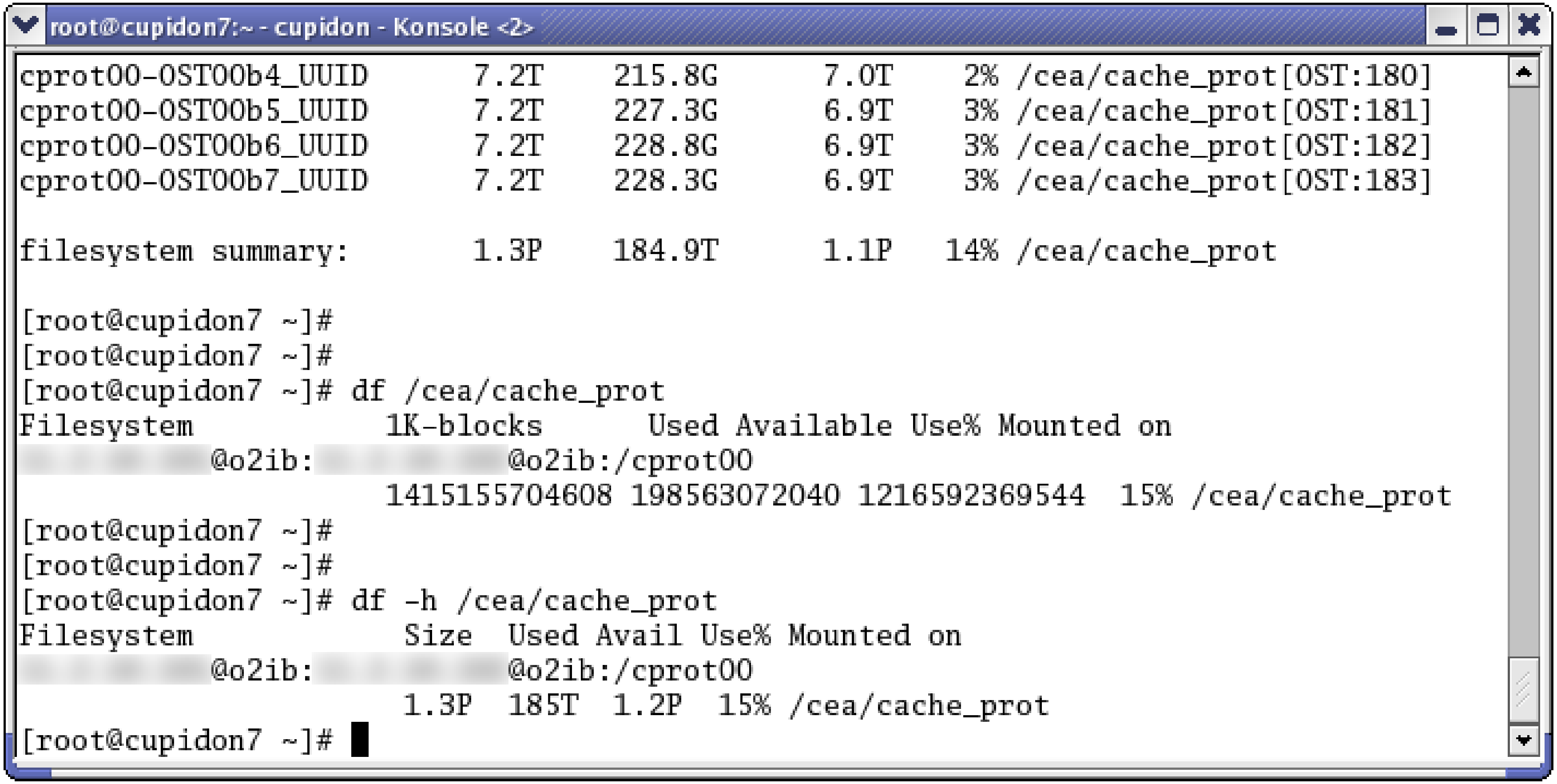
fiber link

Copper link



Capacity File System

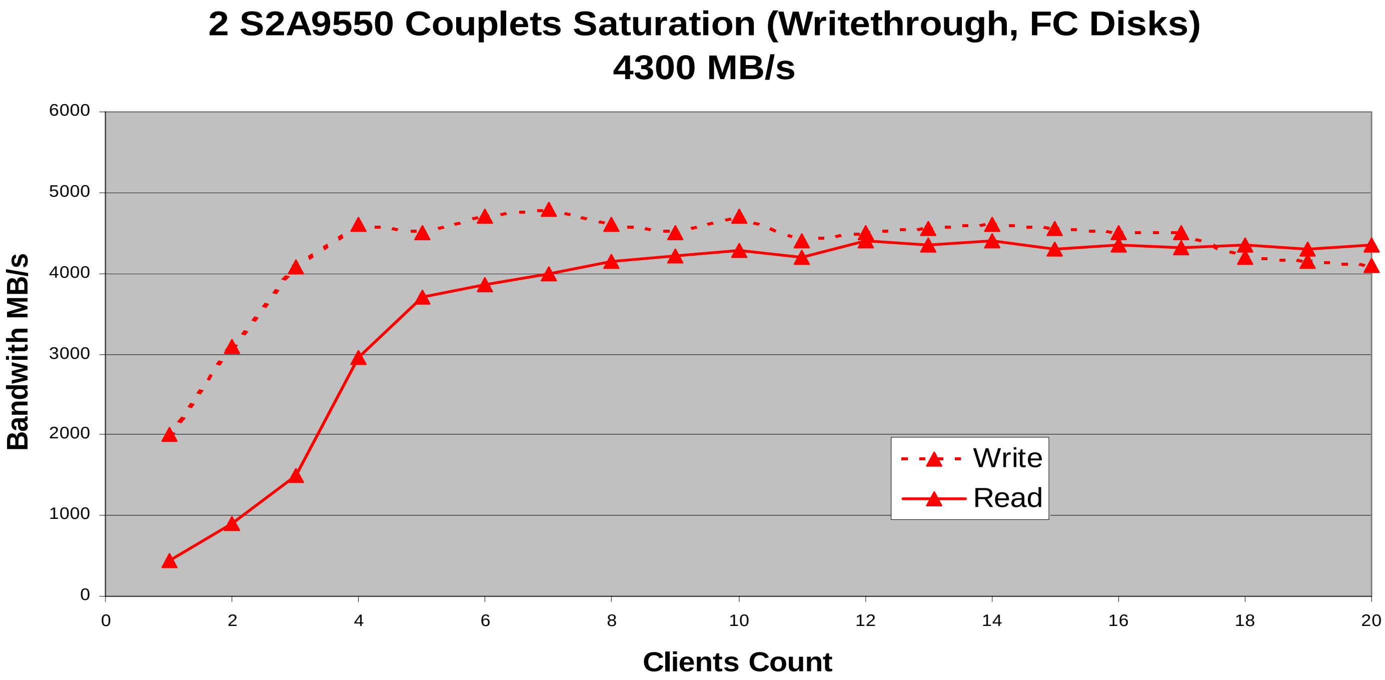
- **Initially created with 2 DDN couplet = 2 * 330 TB**
- **Few weeks ago extended in half a day with 2 DDN**
  - No need to reformat
  - Now at 1.3 PB in one FS



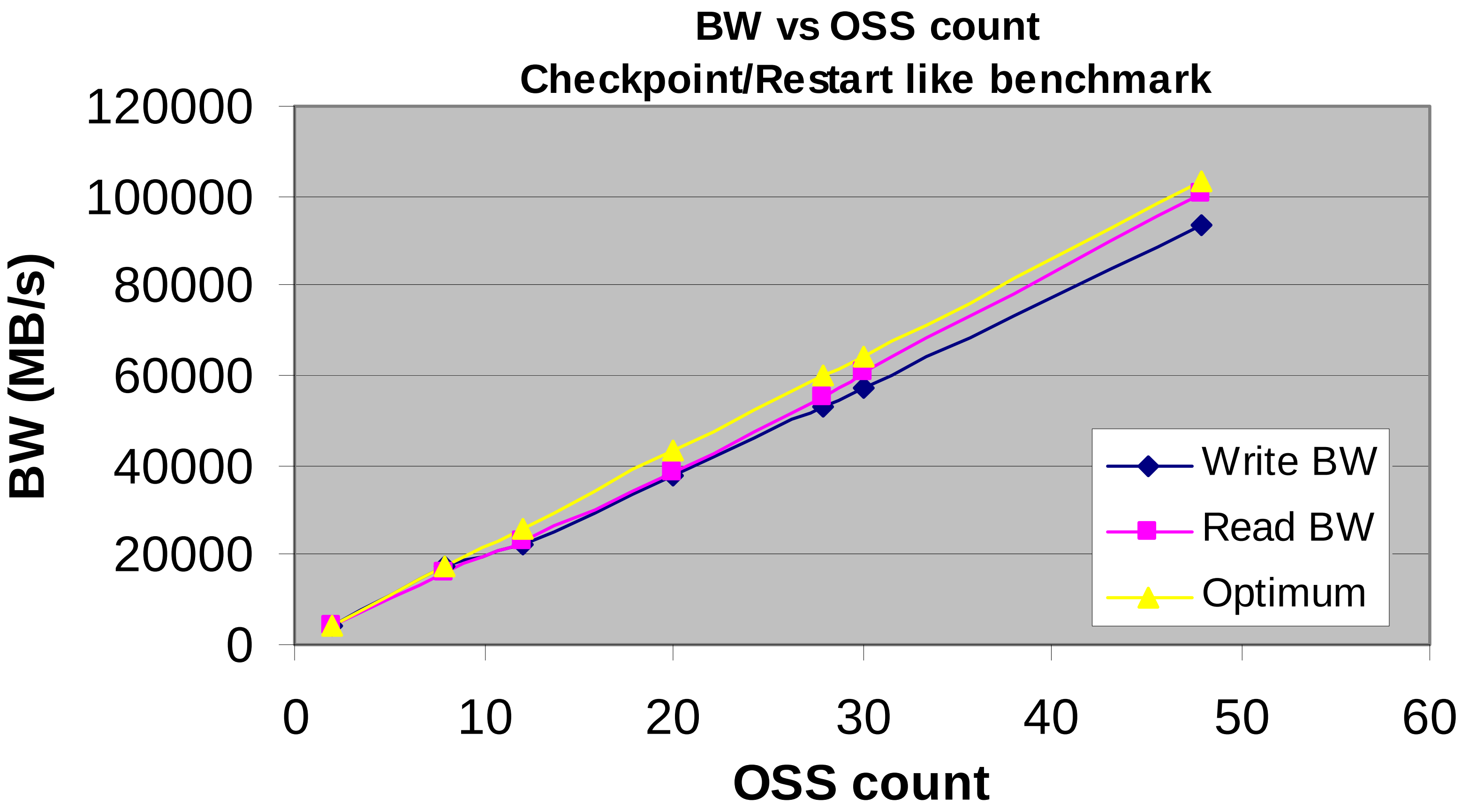
```
cprot00-OST00b4_UUID      7.2T    215.8G      7.0T    2% /cea/cache_prot[OST:180]
cprot00-OST00b5_UUID      7.2T    227.3G      6.9T    3% /cea/cache_prot[OST:181]
cprot00-OST00b6_UUID      7.2T    228.8G      6.9T    3% /cea/cache_prot[OST:182]
cprot00-OST00b7_UUID      7.2T    228.3G      6.9T    3% /cea/cache_prot[OST:183]

filesystem summary:       1.3P    184.9T      1.1P   14% /cea/cache_prot

[root@cupidon7 ~]#
[root@cupidon7 ~]#
[root@cupidon7 ~]# df /cea/cache_prot
Filesystem           1K-blocks      Used Available Use% Mounted on
[redacted]@o2ib:[redacted]@o2ib:/cprot00
                     1415155704608 198563072040 1216592369544  15% /cea/cache_prot
[root@cupidon7 ~]#
[root@cupidon7 ~]#
[root@cupidon7 ~]# df -h /cea/cache_prot
Filesystem            Size  Used Avail Use% Mounted on
[redacted]@o2ib:[redacted]@o2ib:/cprot00
                      1.3P  185T  1.2P  15% /cea/cache_prot
[root@cupidon7 ~]#
```

Lustre Performances

- **Metadata**
  - 5000 create or rename / s on one FS
  - 2500 stat / s on one FS
- **IO’s**
  - The global BW increases linearly with the number of clients and saturates



Lustre Scalability on TERA-10



Lustre Administration Feedback

- **Very easy configuration with Lustre 1.6**

- **Mass configuration tools are mandatory for large Lustre site**
  - HP SFS (HP tools, not open source, not free)
  - Bull lustre_utils (Open source)
  - CFS graphical admin tool (not open source, not free) or CFS lustre_config script (in Lustre distrib)

- **Error messages are still for experts**
  - Actions started at CFS to fix this issue

- **Lustre is very robust**

And now ?

- **We continue to grow the Shared filesystem**
- **Start testing LNET routers**
- **Deploy a cross sites Lustre filesystem over a 40 Gb/s WAN network between 4 sites (part of Carriocas project)**
- **We work with CFS to implement HSM features in Lustre and to be able to connect Lustre to an archival system (HPSS initially but will work with any storage server)**

Questions ?