



Lustre at CEA/DAM

Stéphane Thiell, CEA/DAM, France

stephane.thiell@cea.fr

Lustre at CEA/DAM :: Outline



- **CEA Computing Complex**

- Overview
- TERA/CCRT : Applications and Evolution

- **The Lustre filesystem**

- Introduction

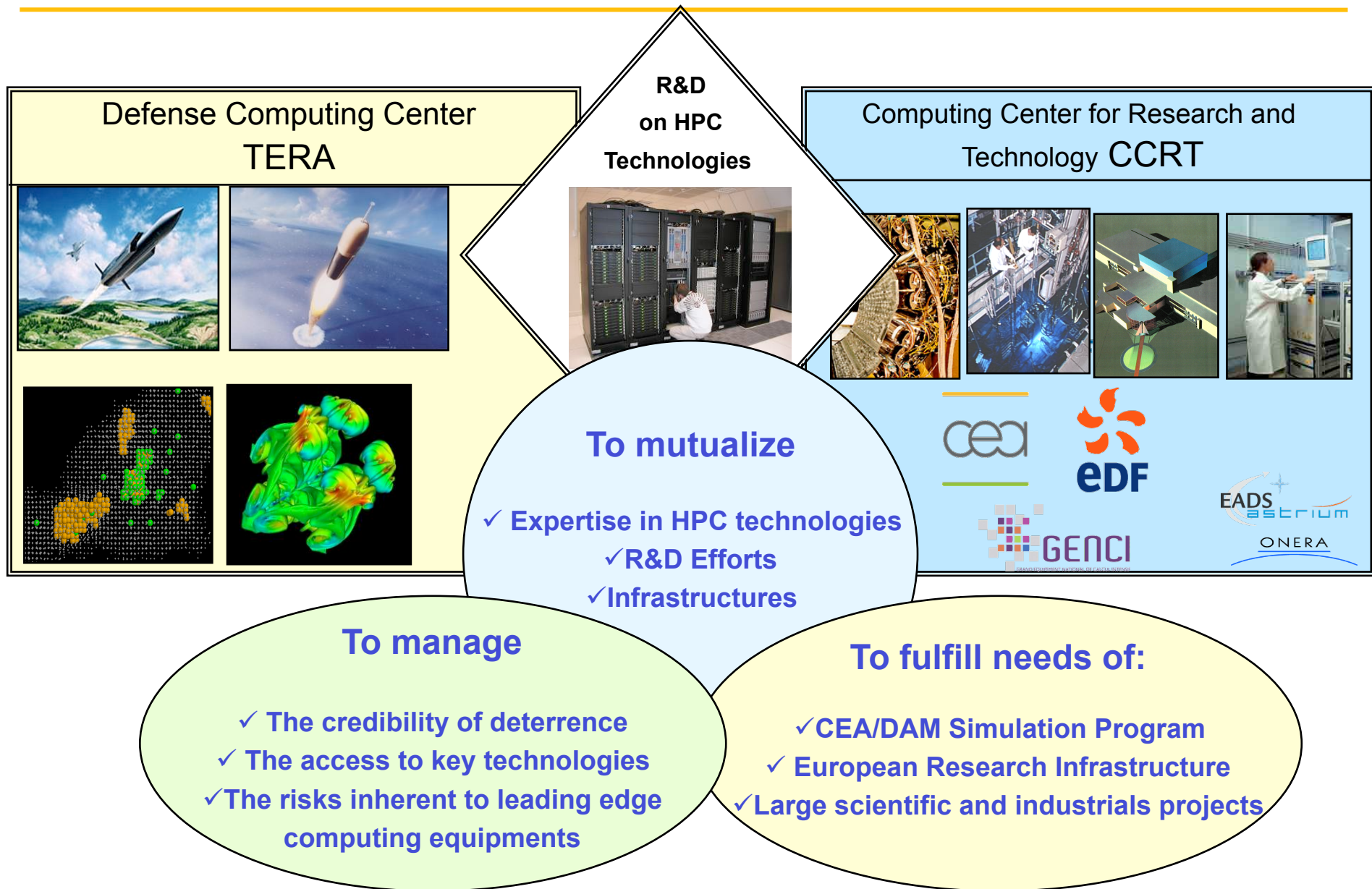
- **TERA-100 and Lustre**

- TERA-100 project overview
- New data-centric computing center architecture
- Servers, Networks and I/O cells configuration
- Lustre Strategy
- Lustre Developments

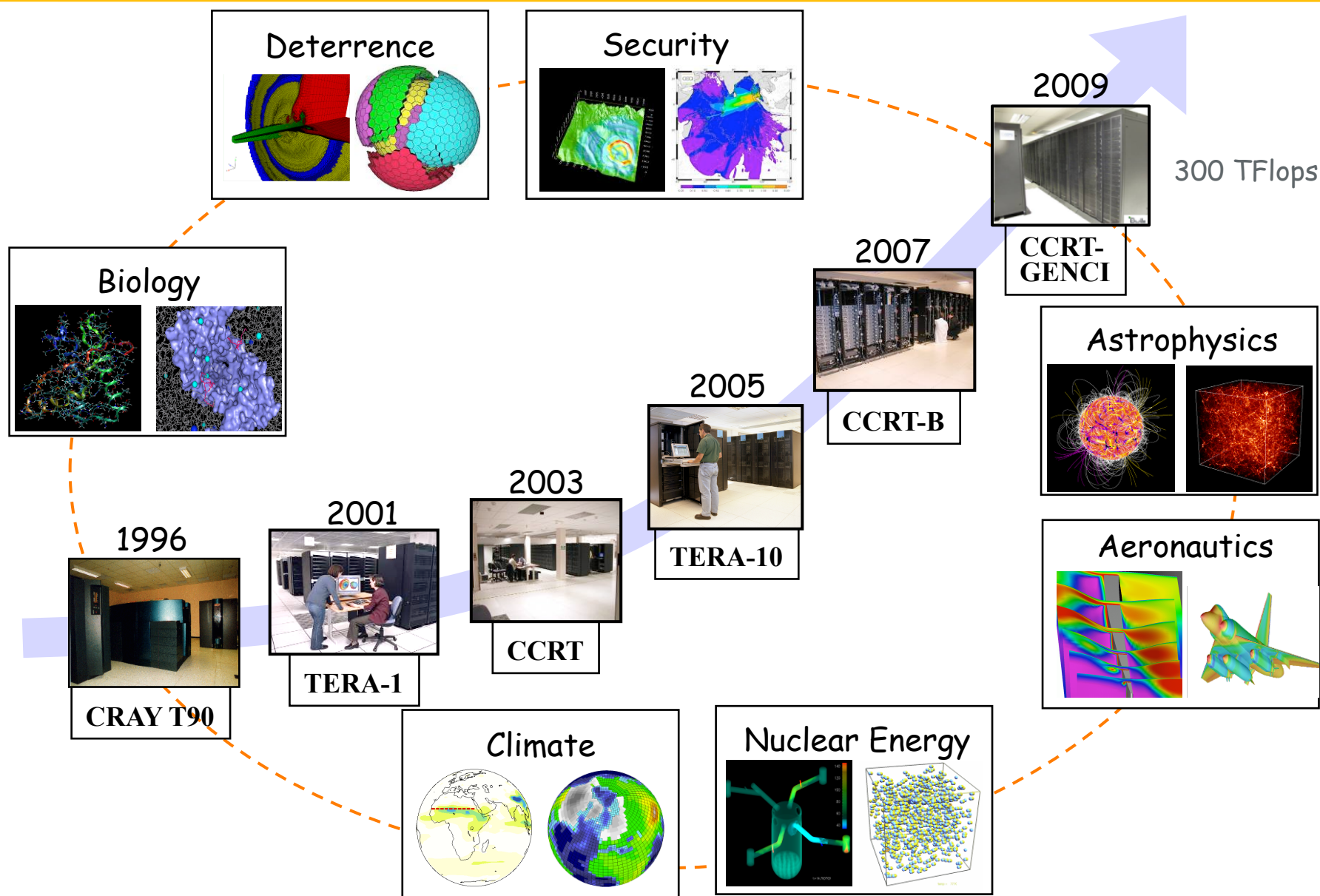
CEA Computing Complex



CEA Computing Complex (cont'd)



TERA/CCRT : Applications and Evolution

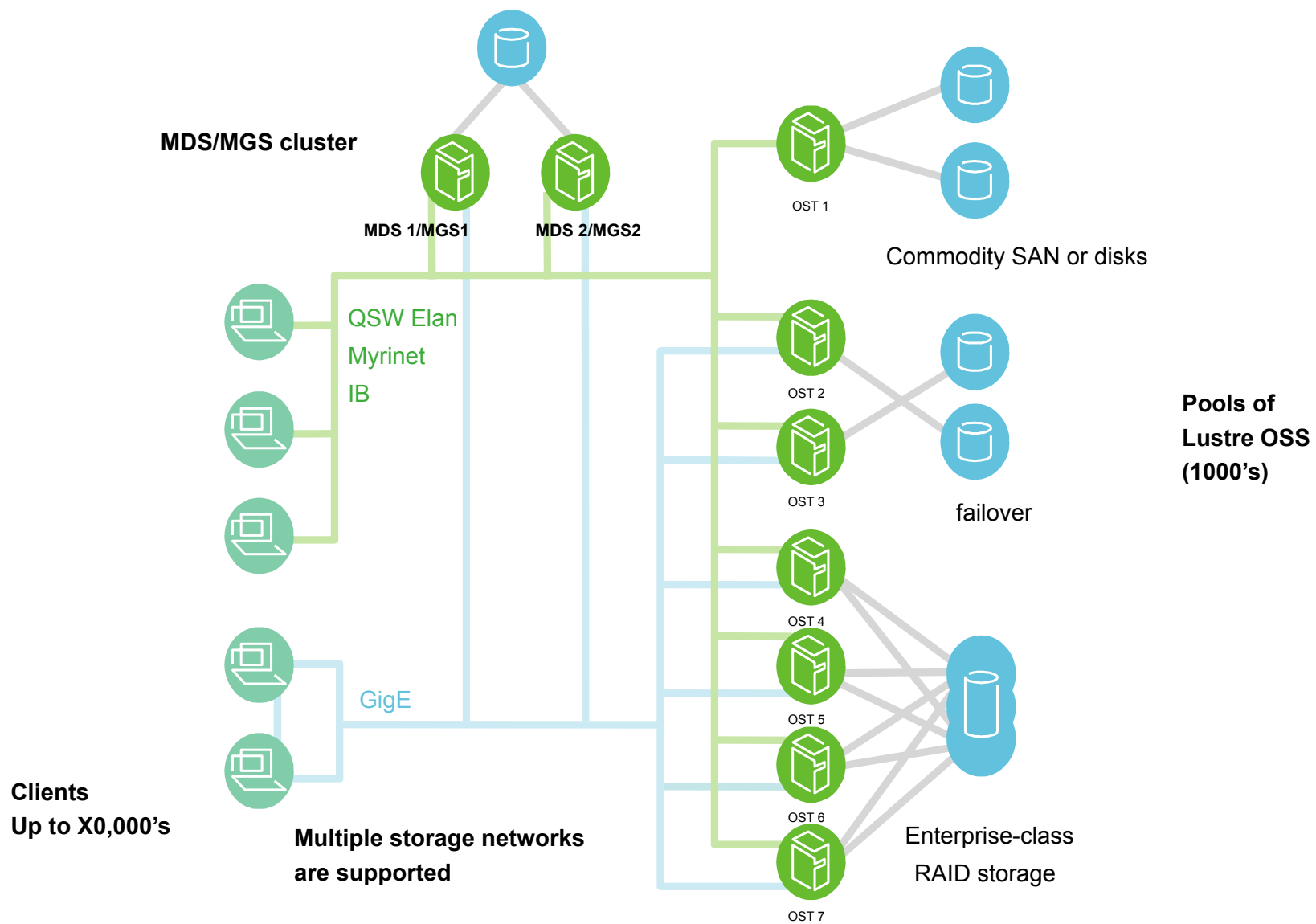


What's Lustre ?



- **A high performance filesystem**
 - A new storage architecture (storage object)
 - Designed for performances
 - ☞ Ten of thousands nodes, peta-bytes of storage, huge directories, ...
 - ☞ 90 % hardware efficiency
- **Open Source Project**
 - Available as tarball and rpm from Sun (RHEL, Suse)
 - ☞ All tools are available to build site specific rpm
 - ☞ RHEL and Suse Linux kernels support (at least)
 - Available through vendors integration (HP, LNXI, Cray, Bull, ...)
- **Managed by Sun as a product, not as a best effort project**

Lustre cluster



Lustre Design Rules



- **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 (eg. ZFS)
 - Uses LNET, a dedicated message passing library
 - ✎ Hardware independence
 - ✎ Transactional RPC or Bulk transfers
 - Support of heterogeneous networks through use of LNET routers
- **IO performances**
 - Large I/O sizes on networks and storage devices
 - Massively parallel
 - Huge client cache
- **Metadata performances**
 - Byte range locking granularity
- **Robust design**
 - High Availability
 - Journals

TERA-100 Project Overview



- **Major goals**

- Multiply by ~20 the classified computing capacity
- Keep one platform, based on COTS
- General purpose (compatible with all programming models)
- Top I/O capabilities for large files

- **Main characteristics**

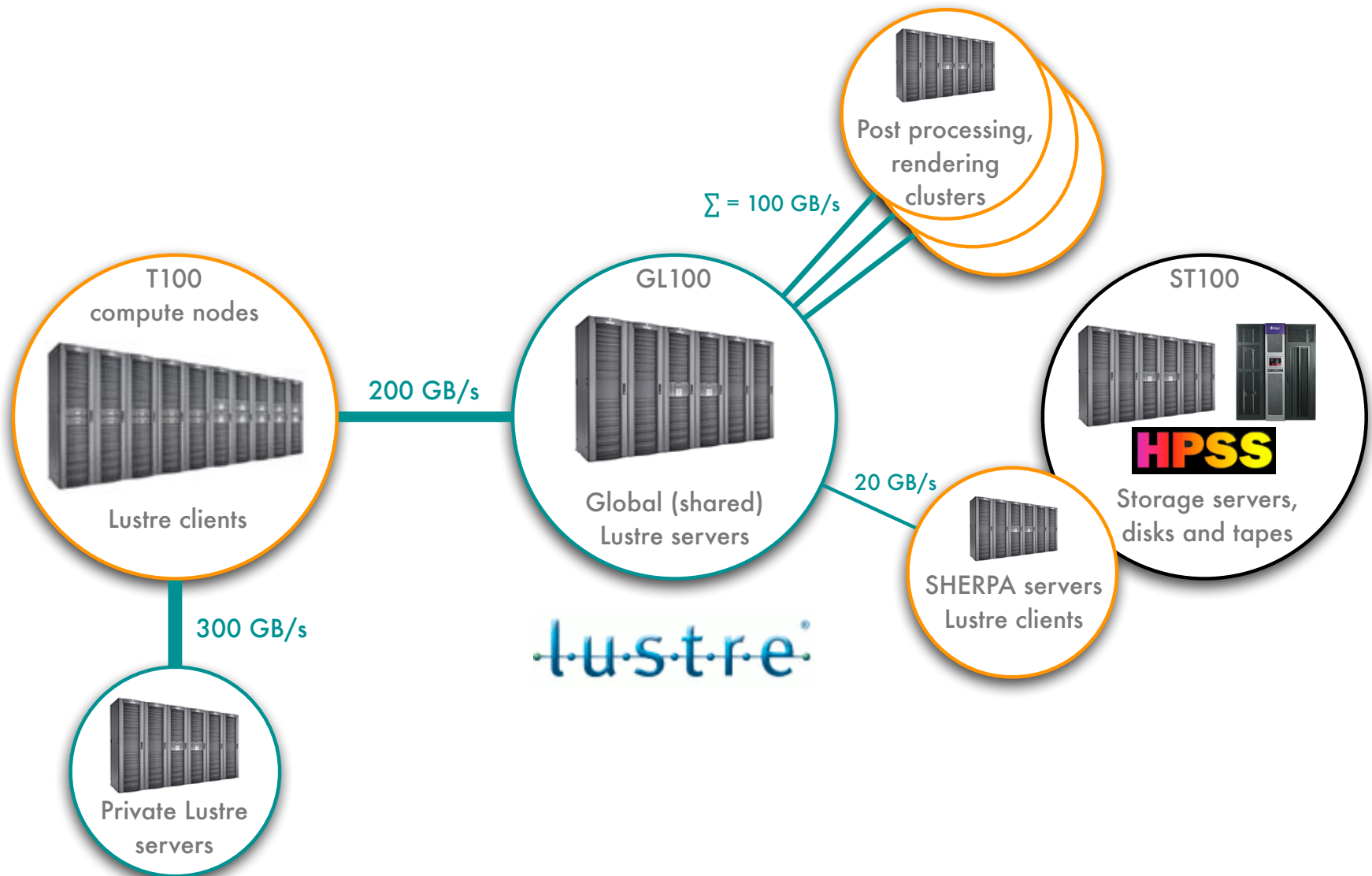
- PetaFLOPS class system
- x86_64 processors
- Some nodes are faster than 500 Gflops
- Up to 2 GB of memory per core
- 300 + 200 GB/s on 2 Lustre filesystems
- Less than 5 MW electric power
- Installation Q2 2010

TERA-100 sizing challenges

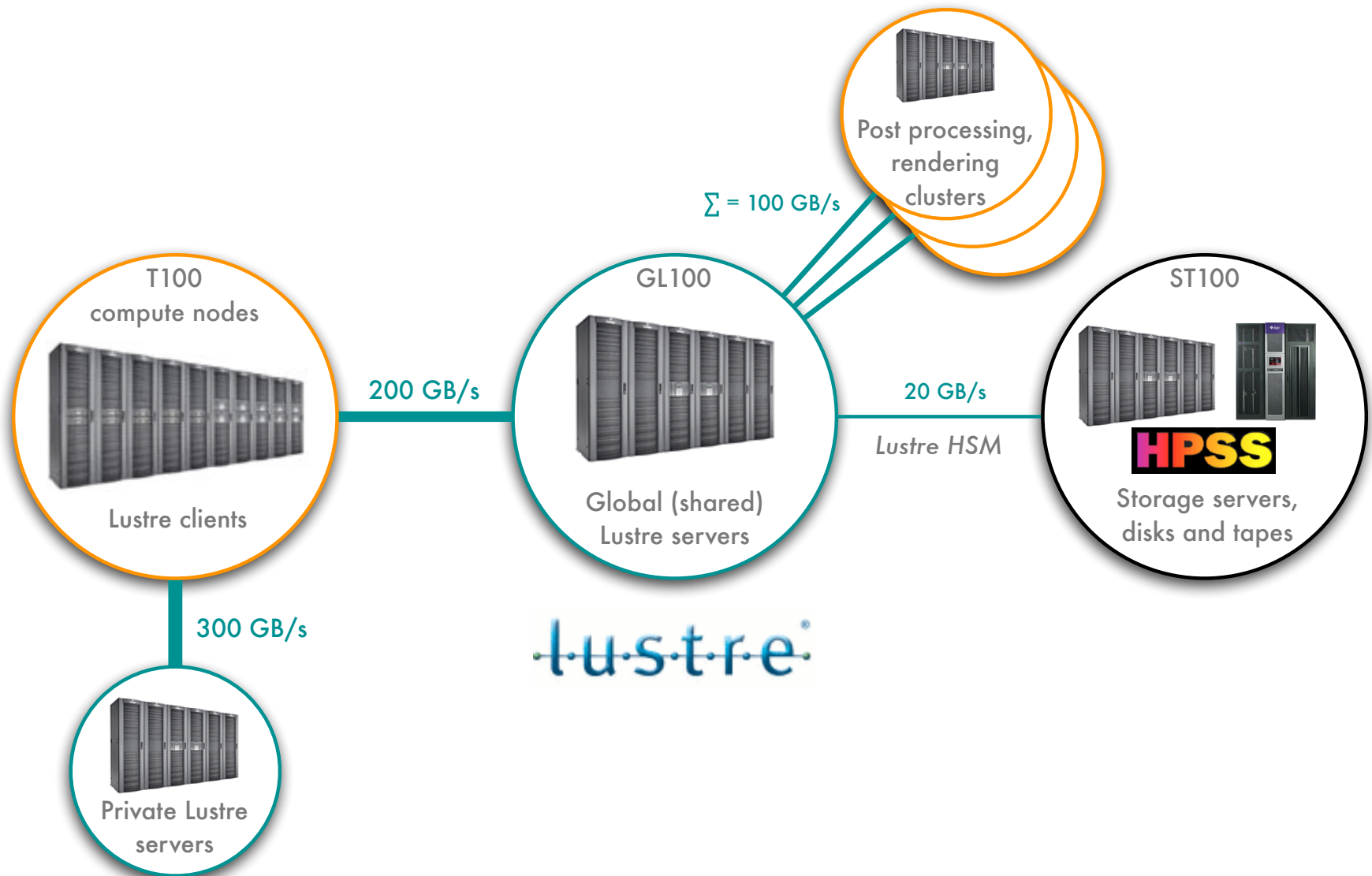


	TERA-1	TERA-10	TERA-100 estim. with same architecture
Peak compute power	5 TFlops	63.8 TFlops	>1 PFlops
Data transfers	5 TB/day	50 TB/day	600 TB to 1 PB/day
Volume growth in HPSS	+1.3 TB/day	+15 TB/day	+200 TB to +300 TB/day

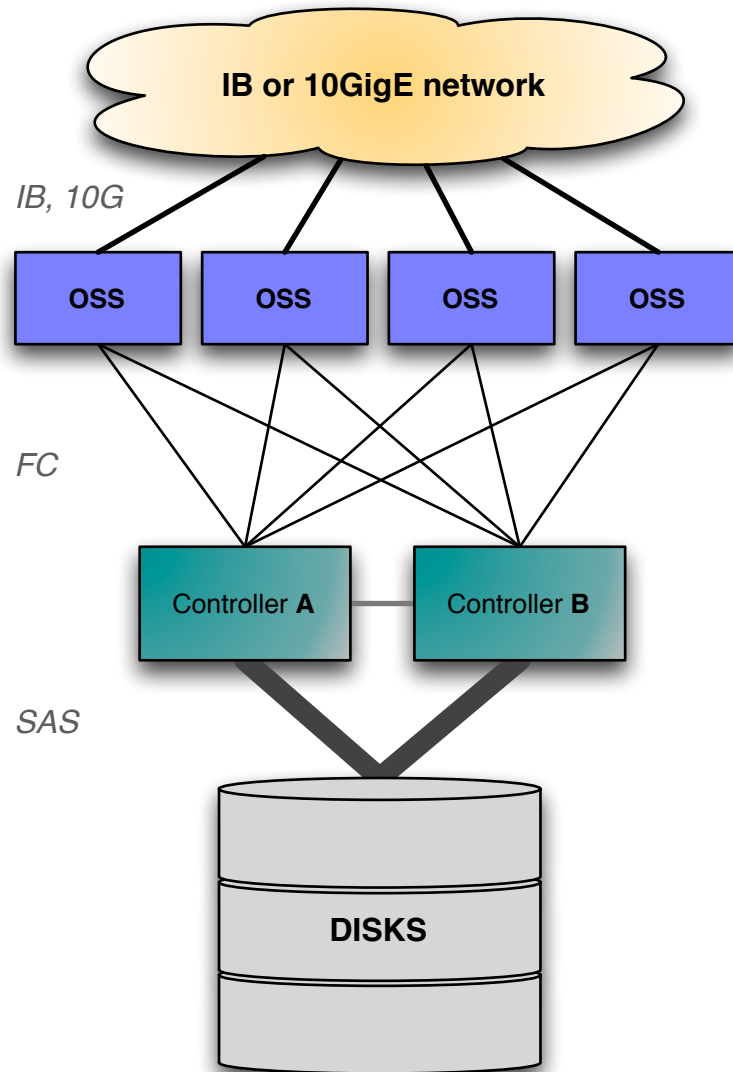
New data-centric compute center architecture



New data-centric compute center architecture (cont'd)

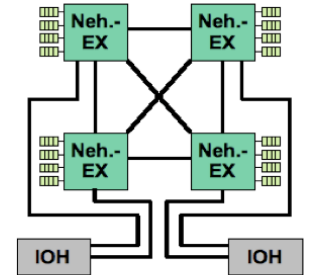


TERA-100 Lustre Servers



- **MDS and OSS nodes**

- 4-socket Nehalem-EX



4 Sockets System

- **4 nodes HA architecture for I/O cells**

- Smooth OST failover

- **Multiple storage systems per I/O cell**

TERA-100 Storage Network

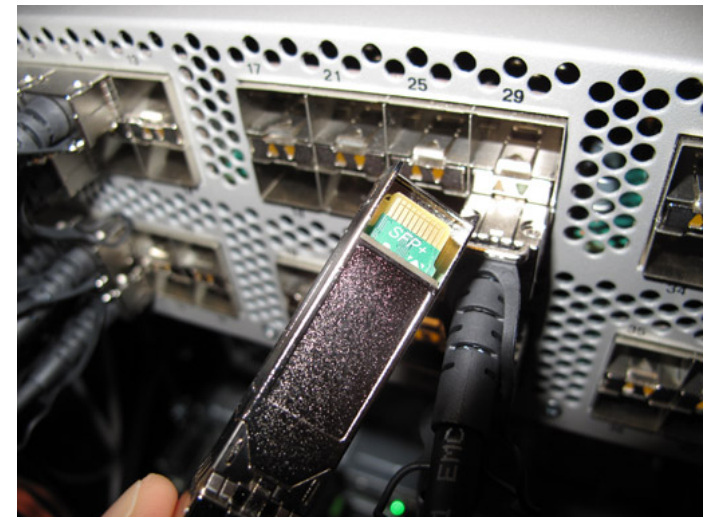


- **2 choices for 200 GB/s**

- InfiniBand QDR

- 10 GbE or 40 GbE (later)

- ➡ With iWARP NICS (Chelsio T3/T4)
 - ➡ LNET tests with iWARP are promising
 - ➡ L2 switches (fully connected)
 - ➡ Waiting for 40 GbE switches



TERA-100 Lustre Strategy



- **Lustre version**

- 1.8 minimum (available today)
- **2.0 targeted (must be available Q2 2010)**

- **LNET**

- Private Lustre File System

- ➡ Native access to InfiniBand QDR network through (**o2ib**)

- Shared Lustre File System

- ➡ Compute nodes access servers through LNET routers
- ➡ Number of routers defines the bandwidth (used for QoS)
- ➡ **tcp** or **o2ib** in iWARP mode for 10 GigE

Manage your data with Robinhood



- **Robinhood is a filesystem accounting and purge tool**

- **Project**

- GPL-compatible, CEA development.
- Website: <http://robinhood.sf.net/>

- **Main features**

- Gather filesystem information from scans (POSIX scan) or event-based (Lustre 2.0).
- Trigger file purges depending on policy rules.
- Trigger file migration depending on policy rules (for Lustre HSM).
- Policies based on file attributes and thresholds.

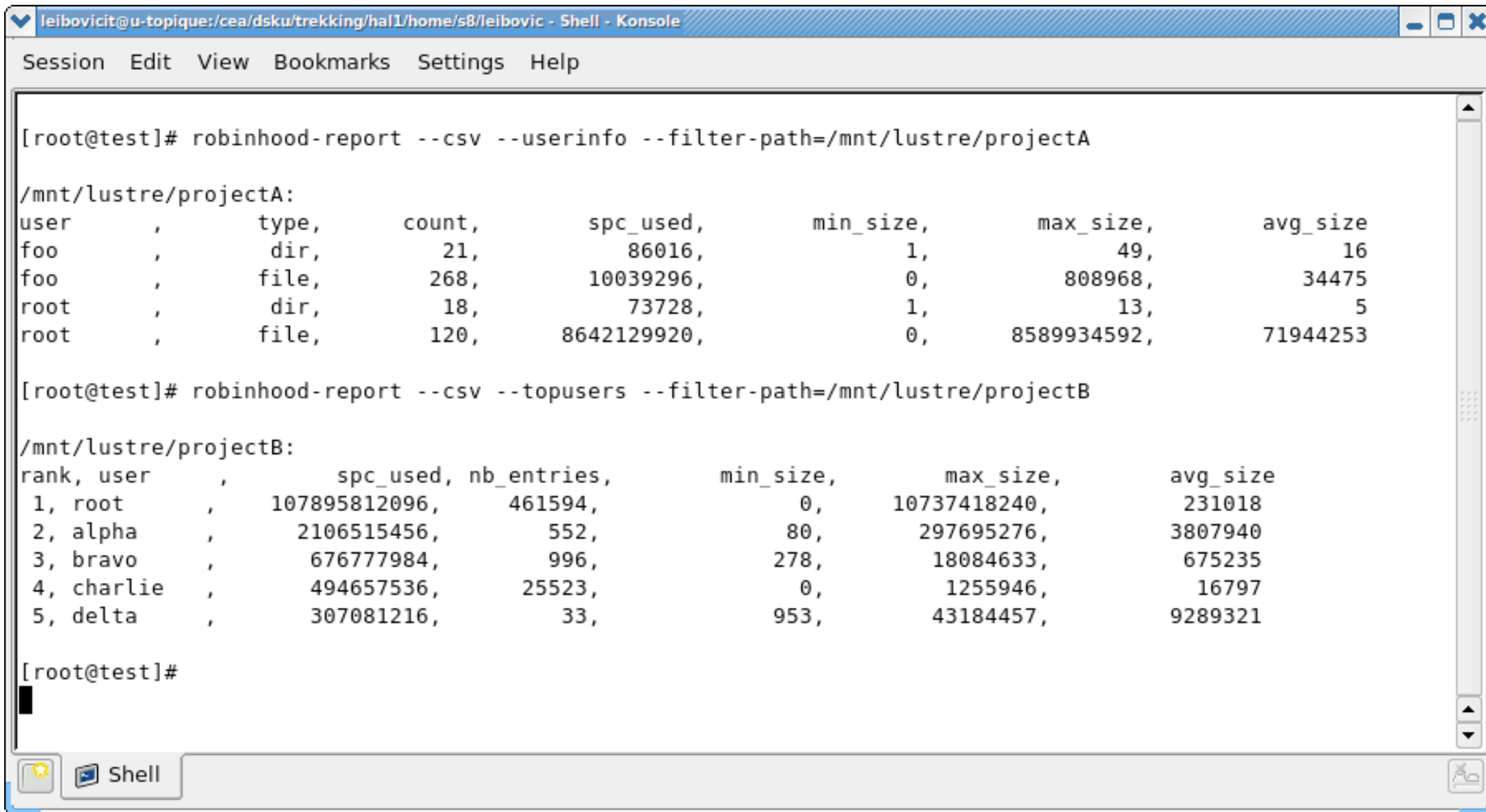
Robinhood example: OST purge

```
leibovic@u-erta:/cea/dsku/trekking/hall/home/s8/leibovic/lustre/demo - Shell - Konsole
Session Edit View Bookmarks Settings Help

# purge OST #1 until its usage decreases to 10%
[root@test ~]# robinhood --purge-ost=1,10
2009/11/04 10:40:24 robinhood@test: ResMonitor | OST #1 usage: 11.90% (55587 blocks) / high watermark: 10.00% (46711 blocks)
2009/11/04 10:40:24 robinhood@test: Released '/mnt/lustre/dir.1/dir.1/dir.2/file.27' using policy 'default',
                                last access 6.0d ago | size=2097152, last_access=1256811816, last_mod=1256811816, storage_units=OST #1
2009/11/04 10:40:24 robinhood@test: Released '/mnt/lustre/dir.1/dir.1/dir.2/file.29' using policy 'default',
                                last access 6.0d ago | size=2097152, last_access=1256811818, last_mod=1256811818, storage_units=OST #1
2009/11/04 10:40:24 robinhood@test: Released '/mnt/lustre/dir.1/dir.1/dir.3/file.5' using policy 'default',
                                last access 6.0d ago | size=2097152, last_access=1256811824, last_mod=1256811824, storage_units=OST #1
2009/11/04 10:40:24 robinhood@test: Released '/mnt/lustre/dir.1/dir.1/dir.3/file.3' using policy 'default',
                                last access 6.0d ago | size=2097152, last_access=1256811822, last_mod=1256811822, storage_units=OST #1
2009/11/04 10:40:24 robinhood@test: Released '/mnt/lustre/dir.1/dir.1/dir.3/file.7' using policy 'default',
                                last access 6.0d ago | size=2097152, last_access=1256811826, last_mod=1256811826, storage_units=OST #1
2009/11/04 10:40:24 robinhood@test: Released '/mnt/lustre/dir.1/dir.1/dir.3/file.9' using policy 'default',
                                last access 6.0d ago | size=2097152, last_access=1256811828, last_mod=1256811828, storage_units=OST #1
2009/11/04 10:40:24 robinhood@test: Released '/mnt/lustre/cthon/file.75' using policy 'empty_files',
                                last access 5.9d ago | size=0, last_access=1256820332, last_mod=1256820332, storage_units=OST #1
2009/11/04 10:40:25 robinhood@test: ResMonitor | OST #1 purge summary: 36864 blocks purged in OST #1 (36864 total)/36860 blocks needed

[root@test ~]#
[root@test ~]#
[root@test ~]#
```

Robinhood example: accounting reports



The screenshot shows a terminal window titled "leibovicit@u-topique:/cea/dsku/trekking/hal1/home/s8/leibovic - Shell - Konsole". The window has a menu bar with "Session", "Edit", "View", "Bookmarks", "Settings", and "Help". The terminal content shows two commands and their outputs.

```
[root@test]# robinhood-report --csv --userinfo --filter-path=/mnt/lustre/projectA
```

/mnt/lustre/projectA:

user	,	type,	count,	spc_used,	min_size,	max_size,	avg_size
foo	,	dir,	21,	86016,	1,	49,	16
foo	,	file,	268,	10039296,	0,	808968,	34475
root	,	dir,	18,	73728,	1,	13,	5
root	,	file,	120,	8642129920,	0,	8589934592,	71944253

```
[root@test]# robinhood-report --csv --topusers --filter-path=/mnt/lustre/projectB
```

/mnt/lustre/projectB:

rank,	user	,	spc_used,	nb_entries,	min_size,	max_size,	avg_size
1,	root	,	107895812096,	461594,	0,	10737418240,	231018
2,	alpha	,	2106515456,	552,	80,	297695276,	3807940
3,	bravo	,	676777984,	996,	278,	18084633,	675235
4,	charlie	,	494657536,	25523,	0,	1255946,	16797
5,	delta	,	307081216,	33,	953,	43184457,	9289321

```
[root@test]#
```

The terminal window also shows a status bar at the bottom with a star icon, a "Shell" button, and a small icon on the right.

Lustre developments for TERA-100: Lustre HSM

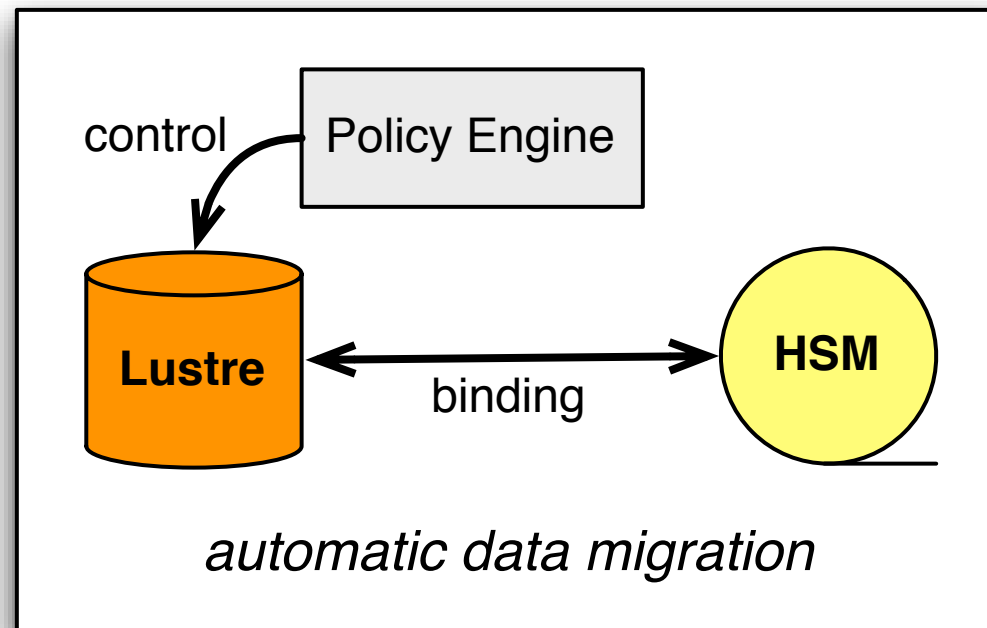


- **Lustre HSM**

- Binding between Lustre and a HSM

- ➡ The goal is to add a transparent frontend to an HSM with the benefits of Lustre performance

- ➡ Will support **HPSS**, **Sun SAM-QFS**, **SGI DMF**



Lustre HSM (cont'd)



- **Lustre HSM (cont'd)**

- Collaboration between Sun and CEA

- Project Status/Timeline:

- ➡ Coding in progress

- ➡ First prototype running

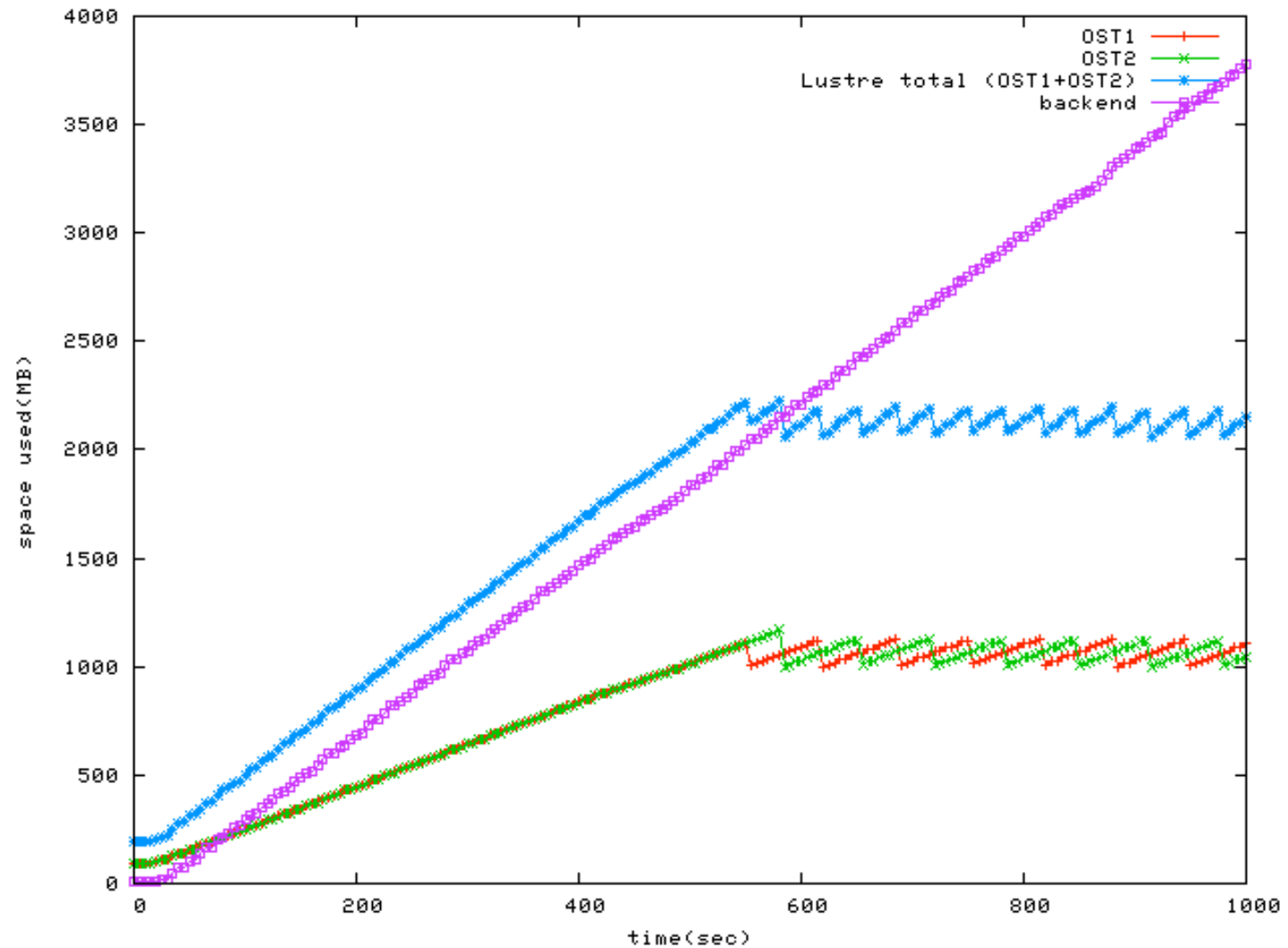
- ➡ Target Lustre 2.x (feature available) and 3.0 (feature supported)

- ➡ First release expected Q4 2009

Lustre HSM (cont'd)



- Lustre HSM first prototype





- **Lustre administration with Shine**

- Common Python library and tool to manage Lustre components in the Compute Center
- Open Source project in collaboration with Bull
<http://lustre-shine.sourceforge.net/>
- Complies with Lustre management features
- Focus on Lustre, eg. tool scalability aspect has moved to a CEA open source project named ClusterShell
<http://clustershell.sourceforge.net/>
- Project Status/Timeline
 - 👉 beta version available today (v0.904)
 - 👉 1.0 in Q4'09 (HA support, fsck, update...)
 - 👉 1.2 in 2010 fully featured: OST Pools, routers, multi-NIDs...



Questions ?