LCE: Lustre at CEA

Stéphane Thiell – CEA/DAM (stephane.thiell@cea.fr)

Lustre at CEA: Outline

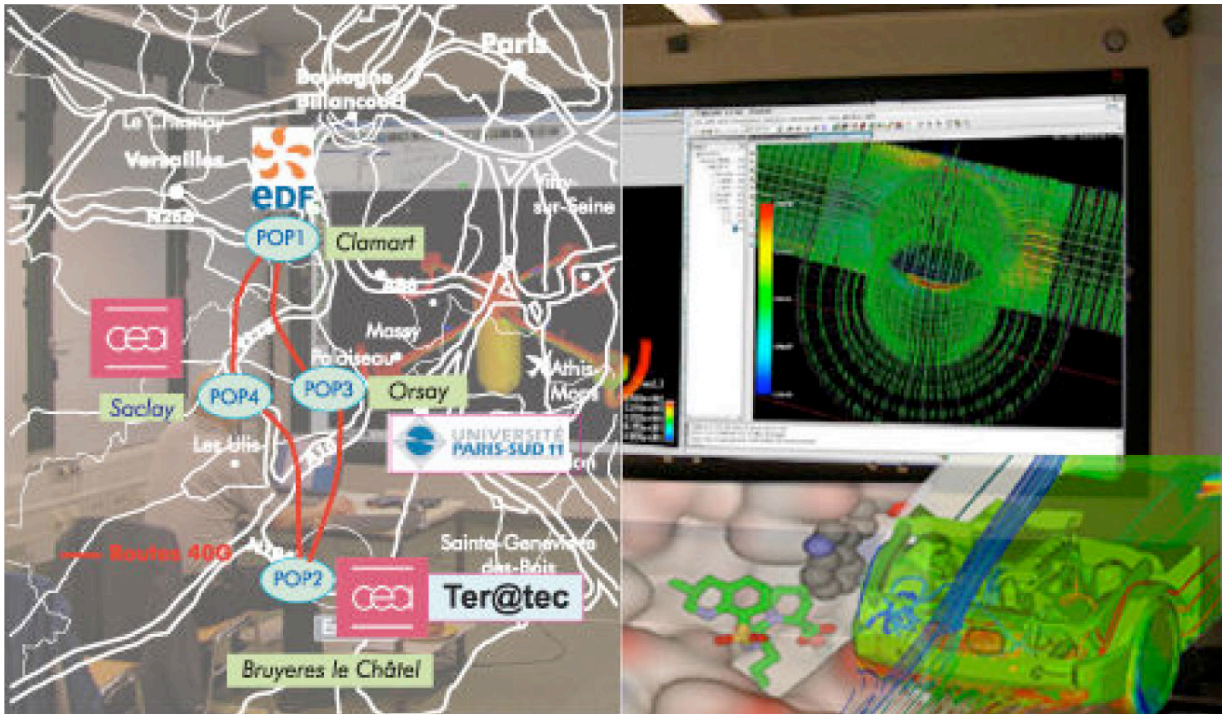
- **Lustre at CEA updates (2009)**
  - Open Computing Center (CCRT) updates
  - CARRIOCAS (Lustre over WAN) project
- **2009-2010 R&D projects around Lustre**
  - Lustre 2.0 early evaluation
  - Hardware: high-end storage systems prototypes
  - Open source projects around Lustre
- **2010: Lustre and the TERA-100 project**
  - Data-centric architecture
  - High-performances, multi-petabytes Lustre filesystems
  - Lustre on TERA-100

Open Computing Center (CCRT) updates

- **2 production Linux compute clusters**
  - Platine: 50 Tflops (IB DDR)
  - Titane: 150 Tflops (IB DDR/QDR mixed)

- **Lustre: plenty of small files**
  - 2 Lustre filesystems per cluster: /scratch and /work
  - 300 TB max per FS
  - Up to 100 millions files seen on /scratch
  - Accounting and monitoring managed by Robinhood

- **Lustre over WAN**
  - 4 sites near Paris in France
    - CEA/DIF Ter@tec, CEA/Saclay, EDF Clamart, Orsay University
  - 40 Gbit/s (one channel) links between sites, 10 Gbit/s NICs
  - One OST pool per site to control files localization



CARRIOCAS: Lustre configuration

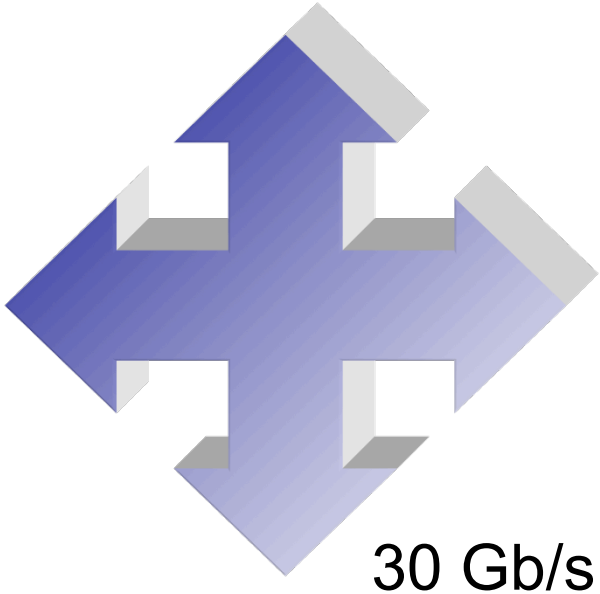
EDF

30 Gb/s

CARRIOCAS

CEA/Saclay

10 Gb/s



10 Gb/s

Orsay/LAL

OST[16-17]

30 Gb/s

OST[18-21]

CEA/DIF

OST[0-15]

MDS

CARRIOCAS project: some results

- **Checkpoint/Restart**
  - 8 Lustre clients at EDF Clamart
  - Servers at CEA/DIF (30 Gb/s max)
  - LNET and TCP/IP tunings needed for WAN
  - Results:
    - 2880 MB/s write (22.4 Gb/s - 75% efficiency)
    - 3120 MB/s read (24.9 Gb/s - 83% efficiency)
- **Remote movie visualization**
  - Visualization wall at EDF Clamart
  - Servers at CEA/DIF
  - Hundred of GB per movie
  - Result: 23,9 millions pixels at 40+ images/sec
    - HD TV: 2 millions pixels at 25 images/sec

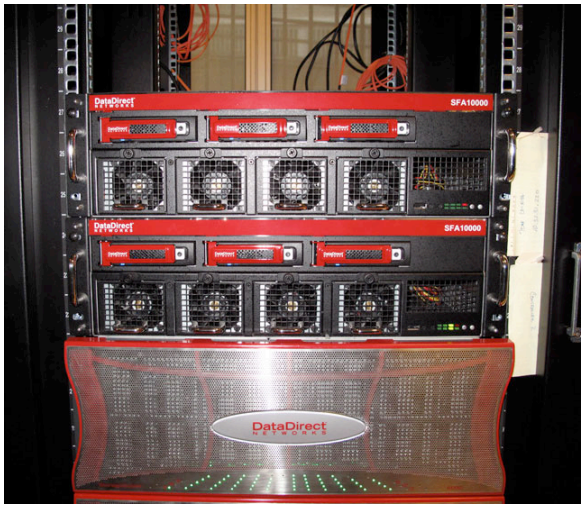
Early Lustre 2 evaluation at CEA/DIF

Some TERA+ Bull R422 nodes

- **Lustre 2.0 on TERA+ cluster**
  - CEA/DAM HPC R&D cluster
  - 8 services nodes, 160 nodes (1280 compute cores)
  - DDN S2A 9550 SATA Lustre storage
  - Bug reporting made easy
- **Lustre 2.0 on TERA-100 *demonstrator***
  - 432 blade nodes cluster (Nehalem-EP)
  - 2 x LSI XBB2 Lustre storage
- **Lustre 2.0 on Global Lustre *demonstrator***
  - DDN S2A 9550 Lustre storage
  - 10 x Sun Fire X4270 Lustre servers
  - Mounted on TERA-100 *demonstrator* through 3 LNET routers

High-end storage technologies prototyping

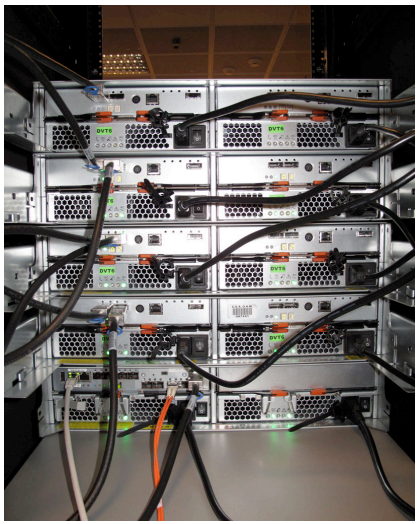
- **DDN SFA10K**
  - Early DDN SFA10K test couplet (spring 2009)
  - Features validation on TERA+ cluster with Lustre 2.0



DDN SFA10K TERA+ test couplet

- **LSI Pikes Peak**
  - Early LSI Pikes Peak SAS2 6 Gb/s prototype controller
  - Multiple SAS 6Gb/s enclosures tests

LSI Pikes Peak SAS 6Gb/s controller (2009 prototype)



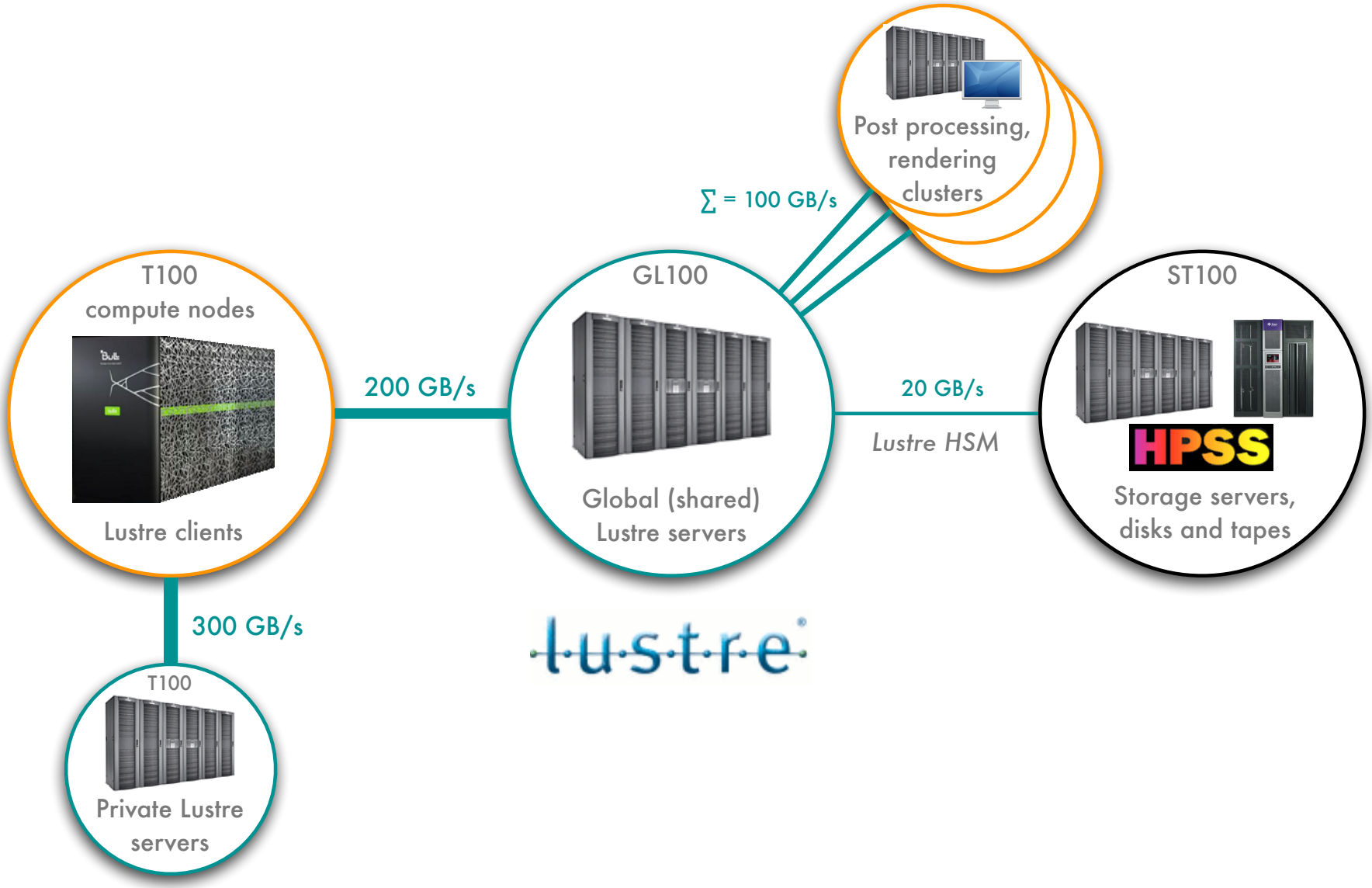
LSI Pikes Peak SAS2 storage system (Camden enclosures)

Open source projects around Lustre at CEA

- **Lustre/HSM binding** – Aurélien's presentation tomorrow morning
- **Shine** – Lustre administration tool (Bull & CEA collab.)
  - Latest version is 0.906, which adds router support and parallel fsck.
  - http://lustre-shine.sourceforge.net
- **Robinhood** – Monitor and purge large filesystems
  - Now supports Lustre 2.0 changelogs
  - http://robinhood.sourceforge.net
- **NFS-Ganesha** – NFS server running in User Space
  - Dedicated backend modules called FSAL (which stands for File System Abstraction Layer) – eg. POSIX, HPSS, …
  - FSAL on top of Lustre 2.0 available since v0.99.52
  - http://nfs-ganesha.sourceforge.net

Lustre on TERA-100

TERA-100 data-centric architecture overview



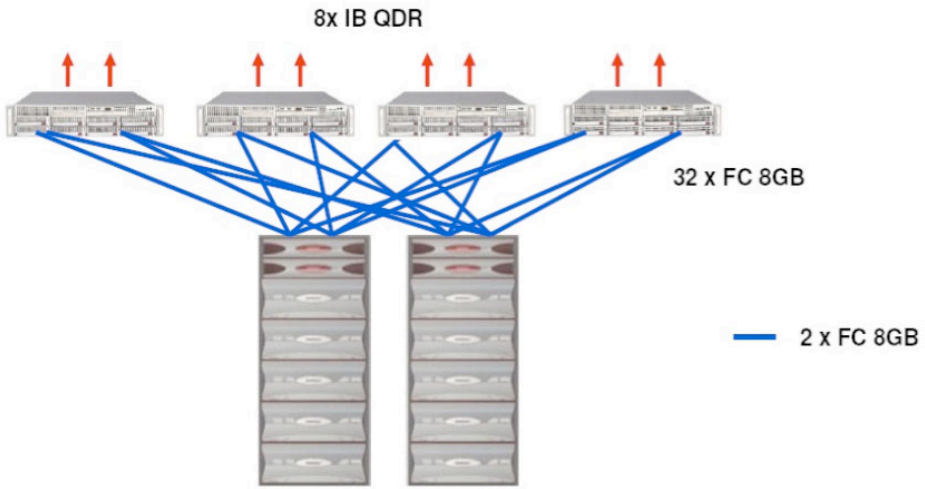
TERA-100 Private Lustre storage

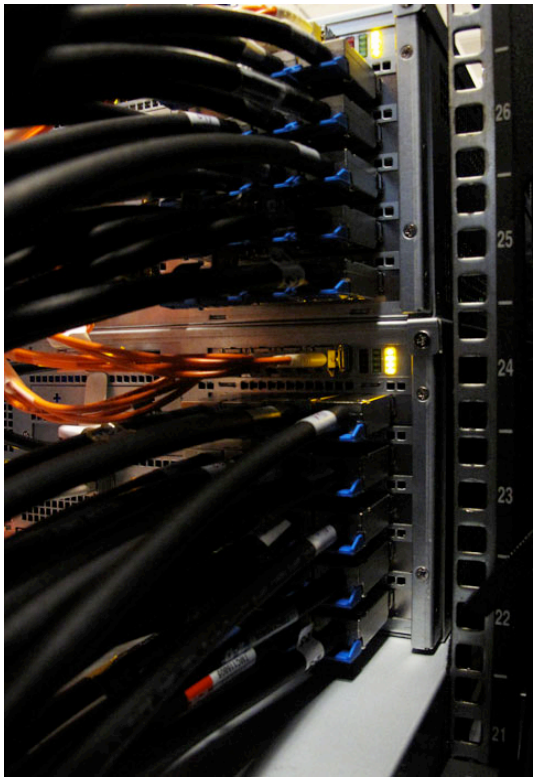
- **Goal**
  - Provide enough bandwidth for checkpoint/restart and temporary files
- **Requirements**
  - **300 GB/s** global bandwidth on Lustre
  - Part of TERA-100 machine (share the cluster interconnect)
  - High disk density
  - Delivery must start Q2'2010 (has started!)

TERA-100 Private Lustre storage architecture

- **Metadata Cell**
  - 4 MDS (Bull Server MESCA 4S3U)
  - 1 DDN SFA10K couplet (300000 IOPS)

- **16 I/O Cells of**
  - 4 OSS (Bull Server MESCA 4S3U)
  - 2 DDN SFA10K couplets (10 GB/s each)





DDN SFA10K SAS 3Gb backend cables

TERA Computing Center Global Lustre storage

- **Data-centric architecture**
  - Zero-copy data access for post-processing clusters
  - Create a very large HPSS cache filesystem

- **Requirements**
  - 200 GB/s bandwidth with TERA-100 (on Lustre)
  - 100 GB/s bandwidth with other clusters
  - Total disk space >15PB
  - High density
  - Delivery by mid-2010

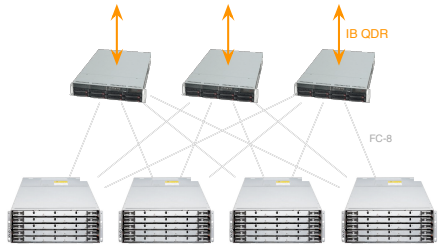
2 choices for Global Lustre storage system (GL100)

- **DDN proposal**
  - Same as Private Lustre Storage (SFA10K)

60 disk slots DDN enclosure (SA6620)

- **LSI proposal**
  - LSI Pikes Peak (with Wembley SAS2 enclosures)

LSI Wembley SAS2 enclosure (60 disk slots)



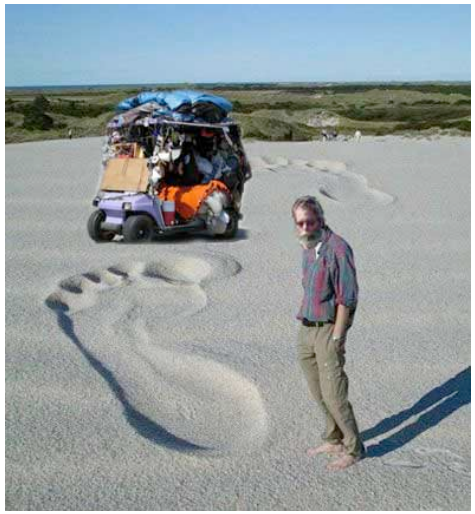
3 nodes IOCell with Pikes Peak

Global Lustre storage network architecture

- **Infiniband QDR storage network**
  - Voltaire QDR Infiniband 4700 switch
- **Lustre routers**
  - 42 LNET routers on TERA-100 (4 x IB QDR each)
  - Networks separation
  - Global filesystem QoS

- **Lustre 2 on TERA-100 and Global Lustre**
  - **Lustre/HSM binding** readiness
  - MDT changelogs (faster Robinhood!)
  - Improved recovery
  - Improved SMP scaling (useful for Bull MESCA nodes)
  - ext4-ldiskfs (larger OSTs)
  - includes Lustre 1.8 interesting features (OST pools, Adaptive timeouts, Version Based Recovery)

- **Lustre administration**
  - Centralized with shine
  - High Availability managed with shine (support smooth OST failover) and Bull tools (based on Pacemaker)



Questions?