Lustre HSM Project

J-Ch Lafoucrière

jc.lafoucriere@cea.fr

- Lustre File System
- Lustre HSM Goals
- Lustre HSM Design

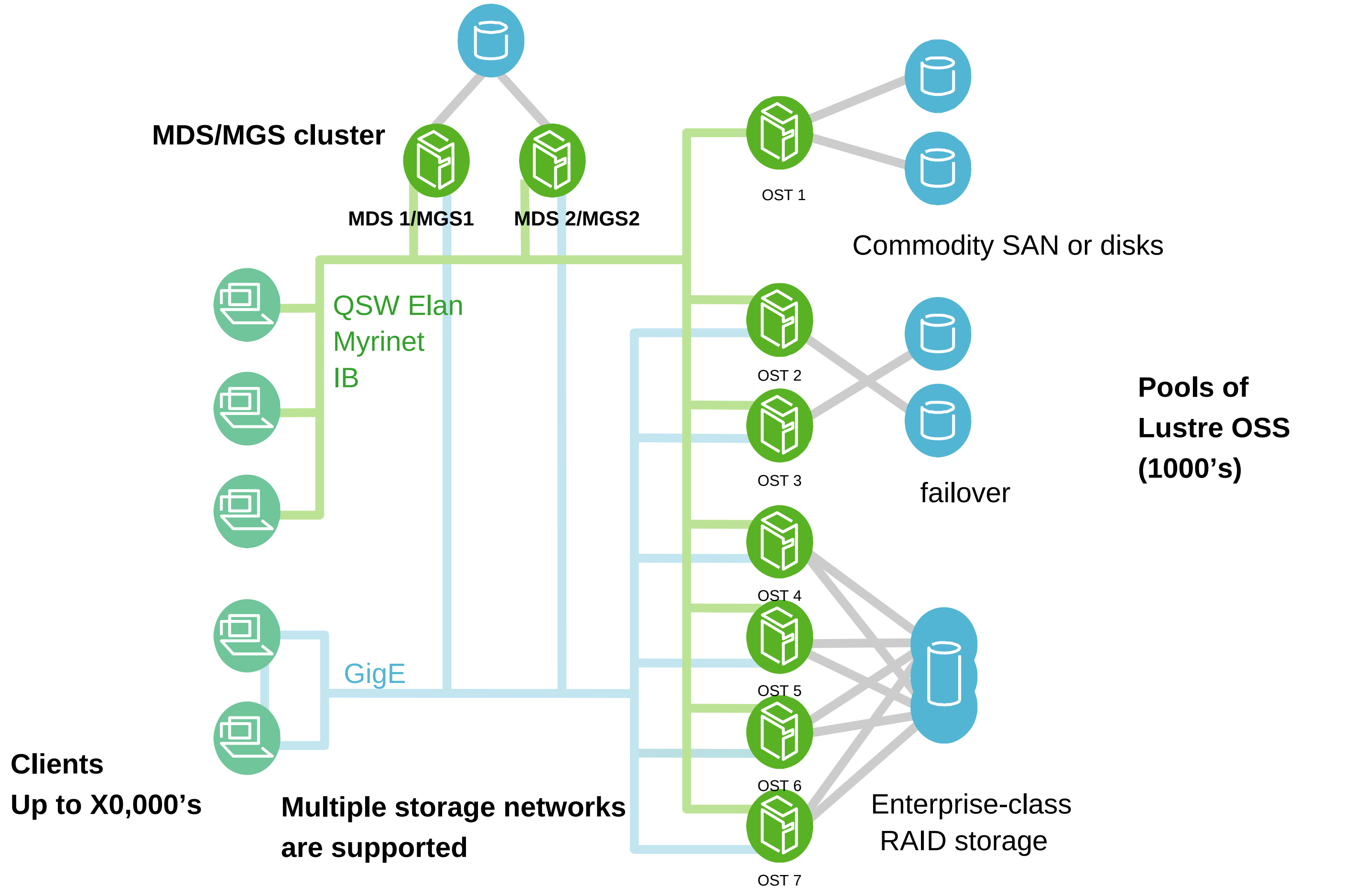
Lustre File System

What’s Lustre ?

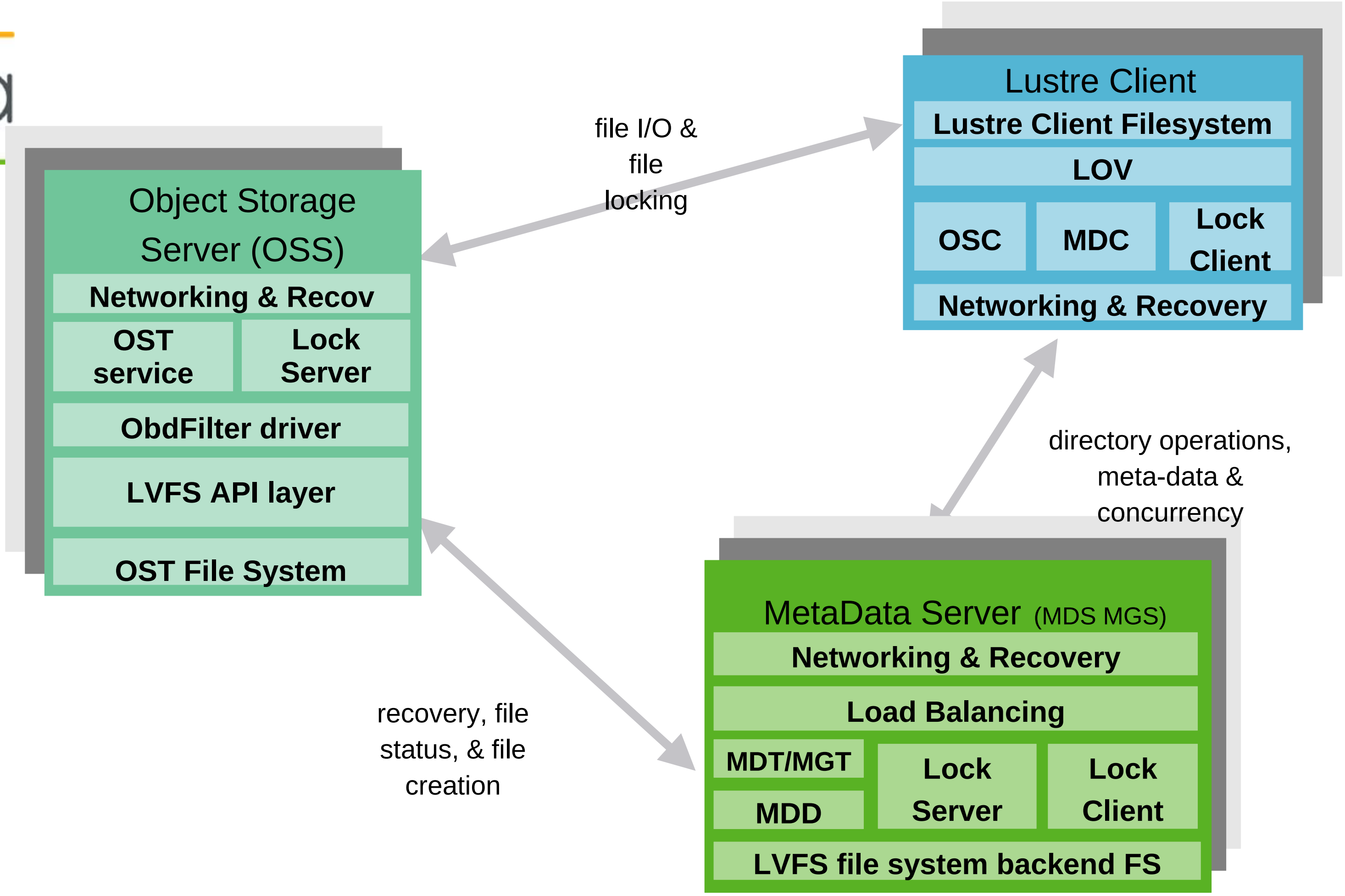
- **A high performance filesystem**
  - A new storage architecture (storage object)
  - Designed for performances
    - X0 000 nodes, Peta bytes of storage, large directories, …
    - 90 % hardware efficiency

- **Open Source Project now at SUN**

Lustre Cluster



Lustre Components



Lustre HSM Goals

Lustre HSM Requirements (1/2)

- **An HSM extension for Lustre**
  - To interoperate with existing storage systems
  - No strong binding with external storage
    - Basic copy-in, copy-out must work with a simple user space tool

- **Provide basic features**
  - Cache miss, archive, purge, transparency
  - Can be used as backup

Lustre HSM Requirements (2/2)

- **All files are always visible in the file system, but a file can reside:**
  - On primary storage (Lustre)
  - On the backend storage
  - On both
- **Metadata (size, …) are always up-to-date**
  - Add a migration status flag
- **Scalable and parallel**
  - Lustre HSM must have a small impact on Lustre performances
  - Target is to impact Lustre performances only when data are not in Lustre (time to bring back data when a cache miss occurs)

Lustre HSM Design

Inside Lustre HSM (1/2)

- **Involve the migration of file system objects**
  - Migration enables multiple Lustre features (HSM, caches for Lustre proxy services, space rebalancing, LAID rebuild, …)

- **Working at a FID granularity level**
  - MDT FID (full file)
  - OST FID (file object)
  - File access by FID feature (obj ID + version)
    - FID is used as the reference key in the backend storage
    - Lustre namespace is independent from backend namespace
  - Unlink in Lustre generates asynchronous unlink in external storage

Inside Lustre HSM (2/2)

- **Use of pre-migration**
  - Automatic
  - On demand: with a user space command
- **File system space management is either:**
  - Automatic
    - At OST level
    - At FS level (MDT)
  - On demand: Based on a provided list of files
- **Purge method**
  - Keep start/end of FID on disk
  - At OST level (objects)
  - At FS level (all file)

Lustre HSM Components (1/2)

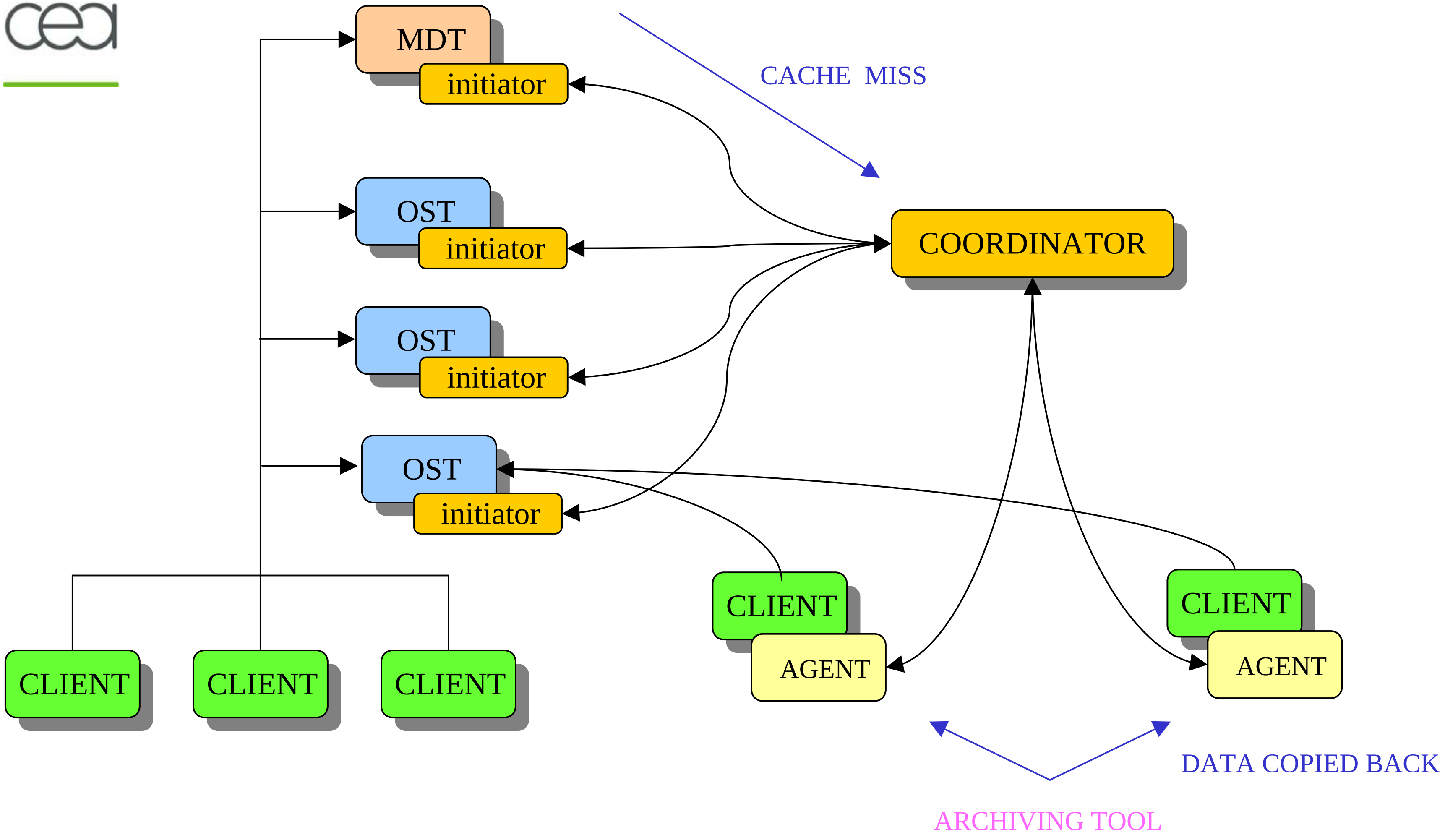
- **Initiators**
  - A node placing a migration request with a coordinating node
  - Handle cache misses
- **Coordinators**
  - A service coordinating migration of data
  - Activate agents to move data
  - Manage multiple requests
  - Send callbacks to initiators
- **Agents**
  - A service used by coordinators to move data, cancel such movement and remove external storage files
  - They invoke HSM tool
- **HSM Tool**
  - A user space tool used to interface to the external storage
  - Copy-in, Copy-out, Remove.

Lustre HSM Components (2/2)

- **Space Manager**
  - A service in charge of pre-migration and space management
  - Use of migration policies

- **Scanners**
  - A tool used to generate list of files without going through the namespace
  - Depend of the MDT backend

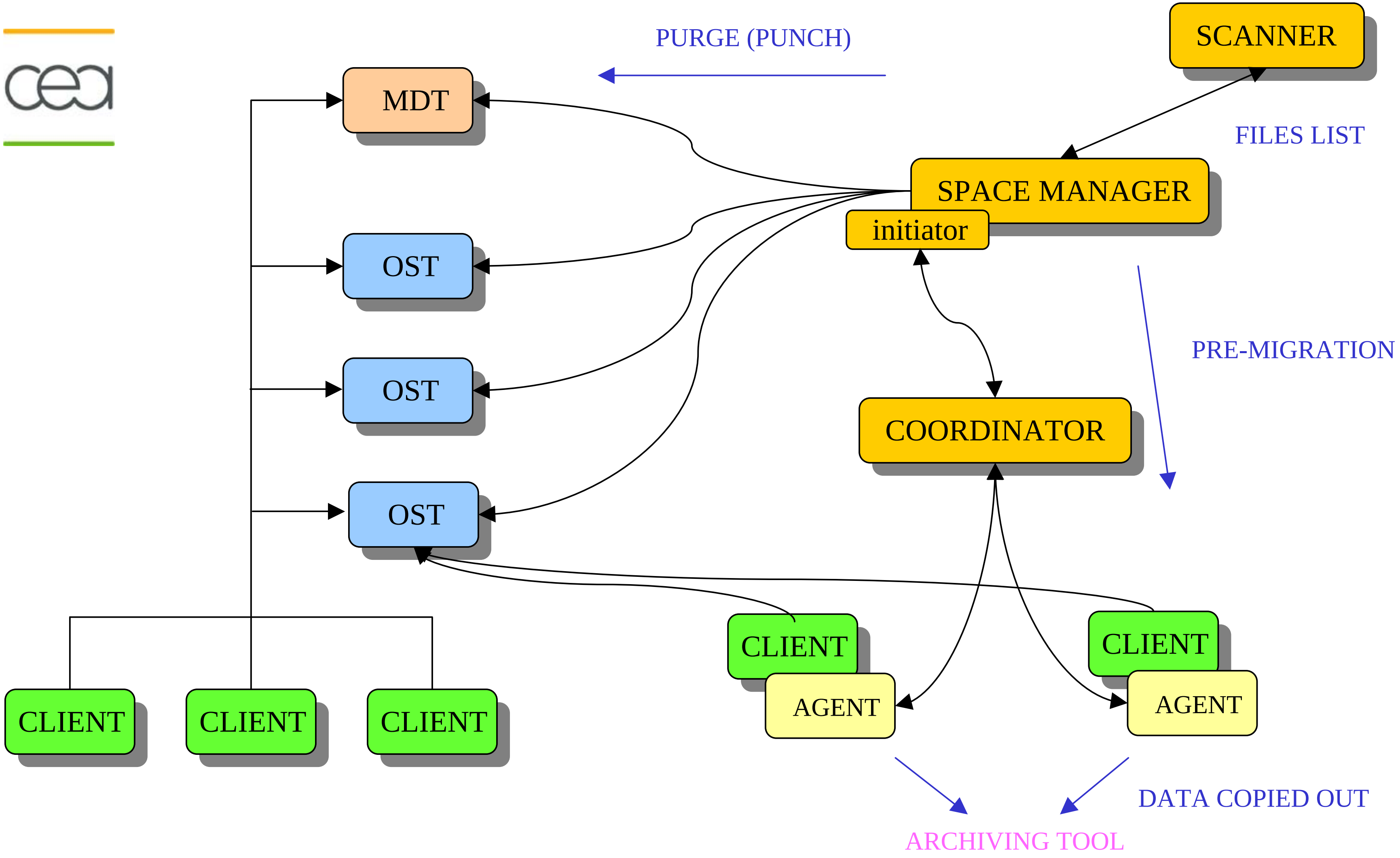
Migration Architecture



External HSM requirement

- **A userspace command able to**
  - Copy from posix (Lustre) to HSM
  - Copy from HSM to posix (Lustre)
  - Remove a file in HSM
  - No Lustre knowledge is needed in the HSM
  - Manage a data transfer cursor
- **HSM namespace based on Lustre FID**
- **A reference to HSM object ID and a version number (returned by HSM) is kept in Lustre**
- **Support of Named Attributes in HSM will allow**
  - Backup of file name in HSM (at migration time)
  - Backup of some file attributes in HSM (at migration time)

Space Management Architecture



Project Status

- **Project is a collaboration with SUN**
  - Architecture design was made by Lustre designers and CEA
  - Coding is made by CEA
- **Lustre target is 1.8.X or 2.0**
- **Architecture done**
- **High Level Design Documents: January 2008**
  - Describe all the components API
- **Detailed Level Design Documents: March 2008**
  - Pseudo Code
- **Code: Summer 2008**
  - HPSS copy tool already made at CEA

Questions ?