LUG 2012 – From Lustre 2.1 to Lustre HSM

Lustre @ IFERC (Rokkasho, Japan)

Diego.Moreno@bull.net

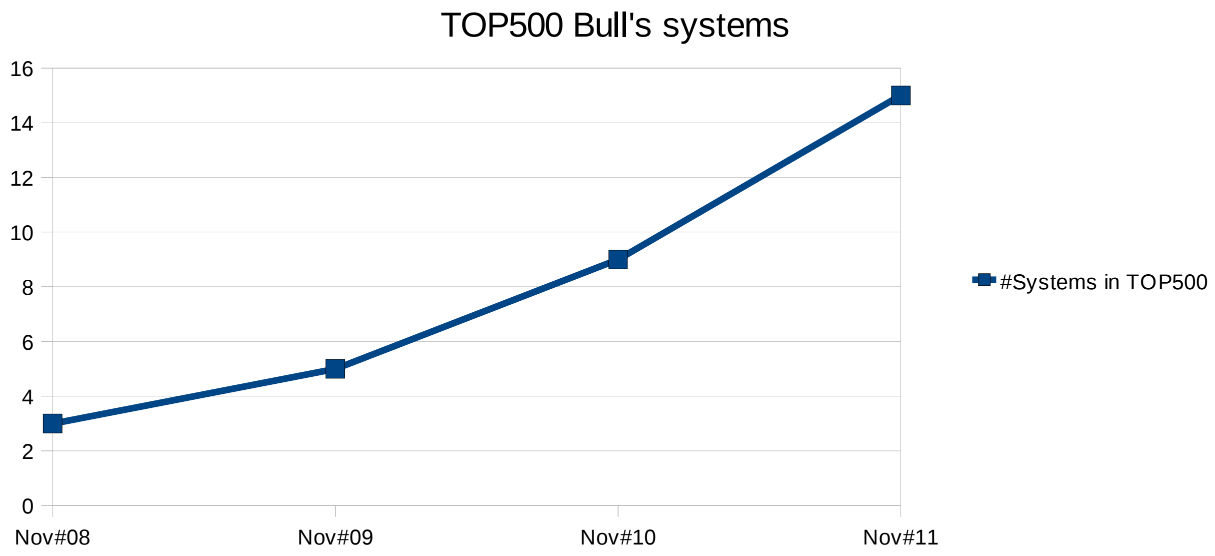
Architect of an Open World™

From Lustre-2.1 to Lustre-HSM - Outline

- About Bull
- HELIOS @ IFERC (Rokkasho, Japan)
- Lustre-HSM
  - Basis of Lustre-HSM
  - HSM patches and new layout lock
  - Testing Lustre-HSM
  - Lustre-HSM Roadmap
- Lustre backup
  - Robinhood-backup architecture @ IFERC
  - From robinhood-backup to Lustre-HSM
- Conclusion

About Bull

- In the HPC:
  - 1st European manufacturer
  - 3 petaflop systems in the last 18 months



Bull & Lustre

- In Lustre community:
  - EOFS member
  - First Lustre 2.0 adopter
  - First Lustre 2.1 installation in a petaflop machine
  - Other lustre contributions:
    - NUMIOA architectures
    - Multi-attachment infiniband configuration
    - Multipath tuning patch
    - Redhat 6 kernel adaptations (2.6.32)

EOFS European Open File System

Helios@IFERC (Rokkasho, Japan)

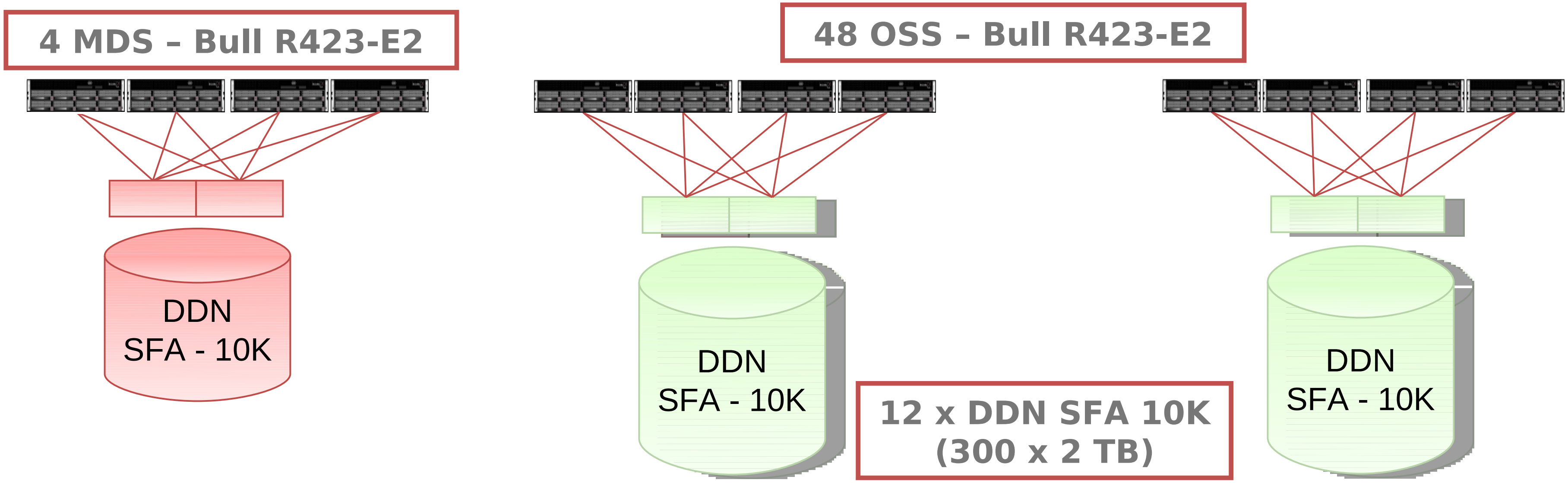
- International Fusion Energy Research Centre
- More than 1.5 Petaflops
- Memory: 280 TB
- 245 bullx® B-chassis:

- **245 bullx B chassis**
- **2205 blades B510**
- **4410 compute nodes**
- **8820 sockets Intel Sandy Bridge 2.7GHz**

Would have been #6 in TOP500 November11

Lustre in Helios

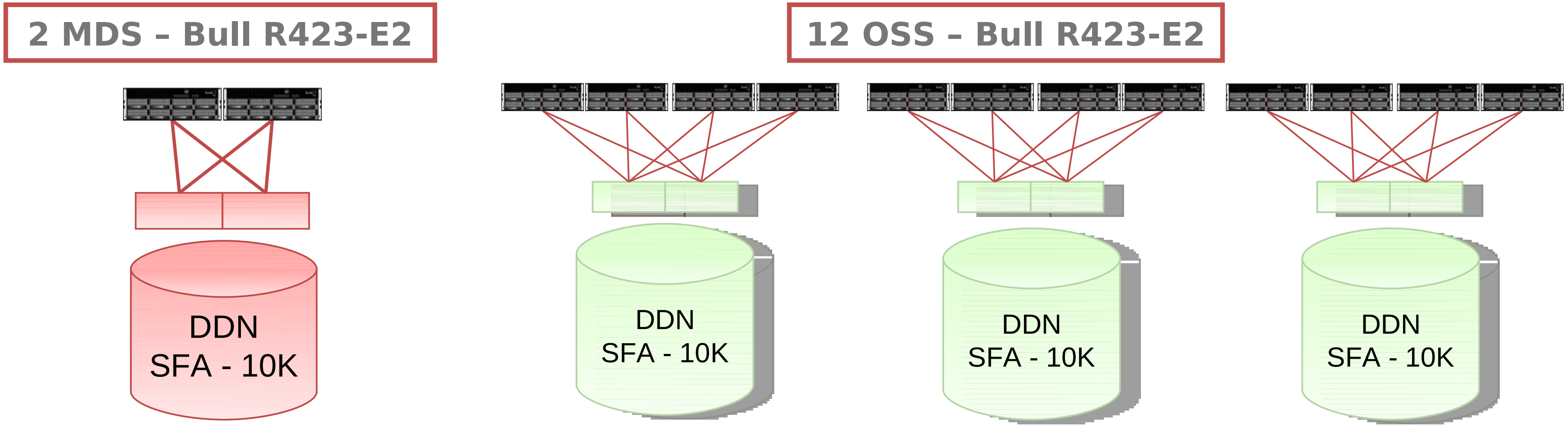
- L1, scratch filesystem at HELIOS:
  - High IO throughput: 110 GB/s
  - Moderate storage capacity: 5.7 PB



Lustre in Helios

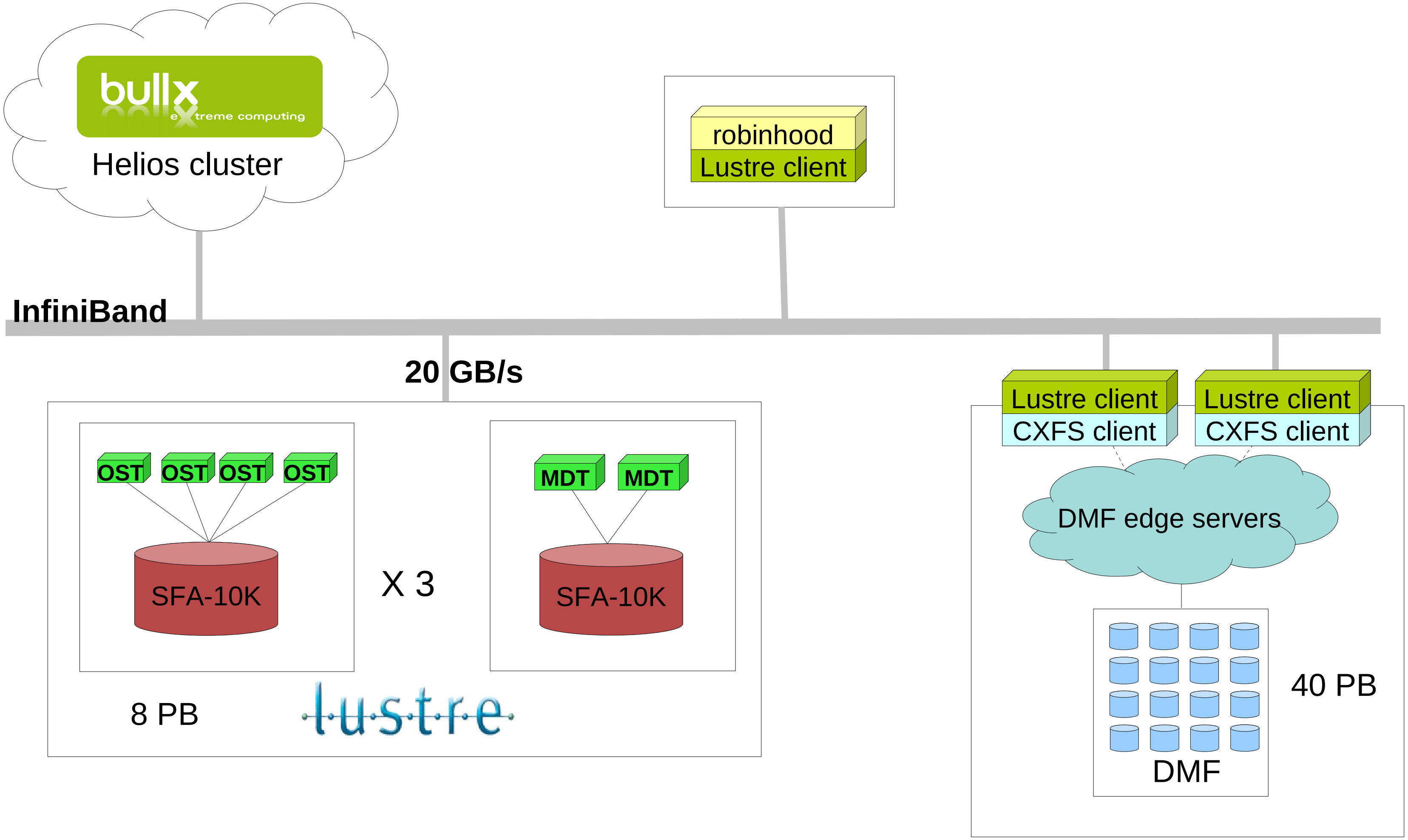
- Second Lustre level, L2:
  - L2, Lustre-DMF filesystem:
    - Moderate IO throughput: 20 GB/s
    - High Lustre capacity: 8.6 PB filesystem
    - Connected to SGI's DMF Edge Servers: 40 PB of extra storage in tape drives

IFERC



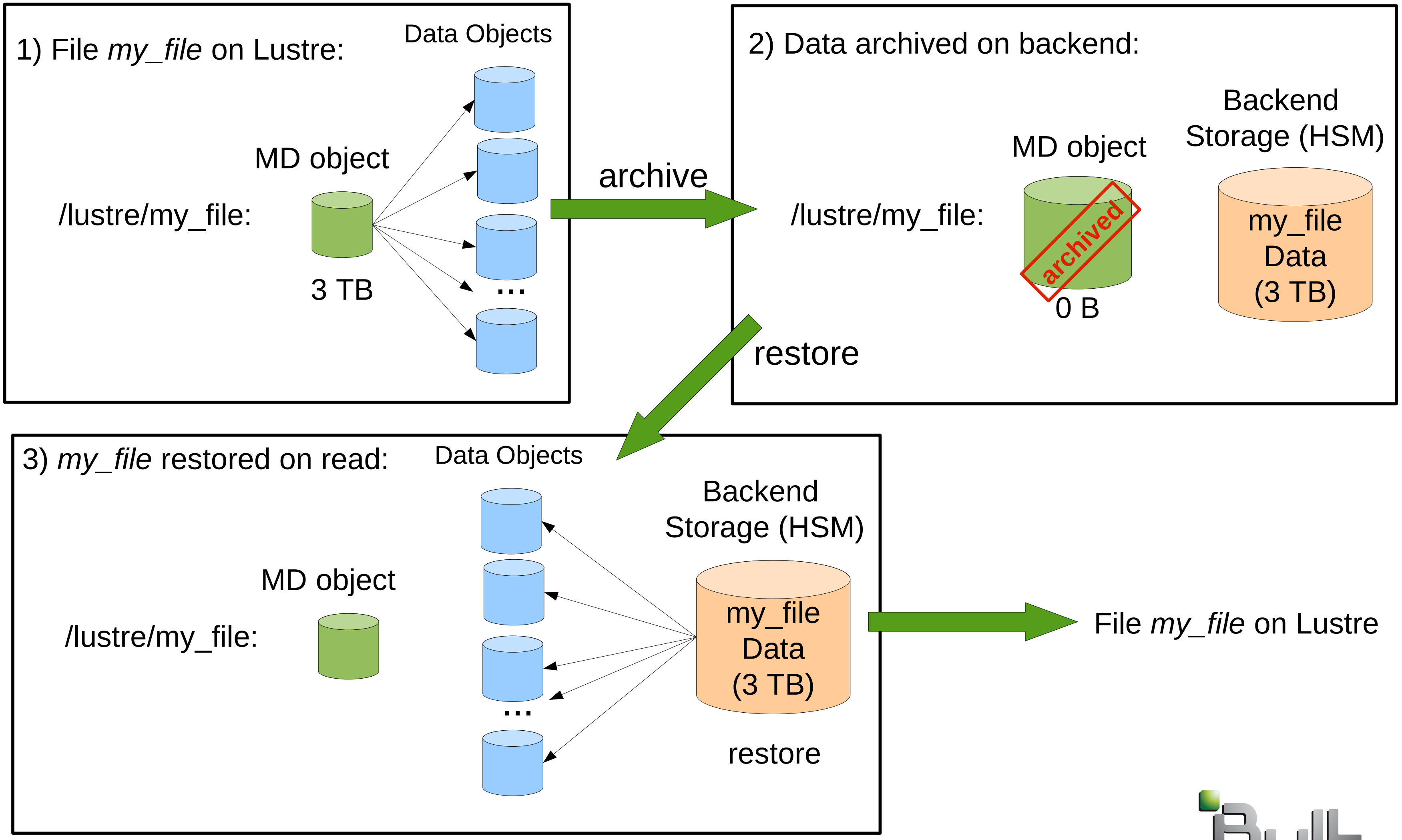
Architect of an Open World™

Archive Lustre L2 – DMF (HSM)



Architect of an Open World™

Lustre–HSM use case



1) File *my_file* on Lustre:

Data Objects

MD object

/lustre/my_file:

3 TB

...

archive

2) Data archived on backend:

MD object

Backend Storage (HSM)

/lustre/my_file:

archived

0 B

my_file Data (3 TB)

restore

3) *my_file* restored on read:

Data Objects

MD object

/lustre/my_file:

Backend Storage (HSM)

my_file Data (3 TB)

...

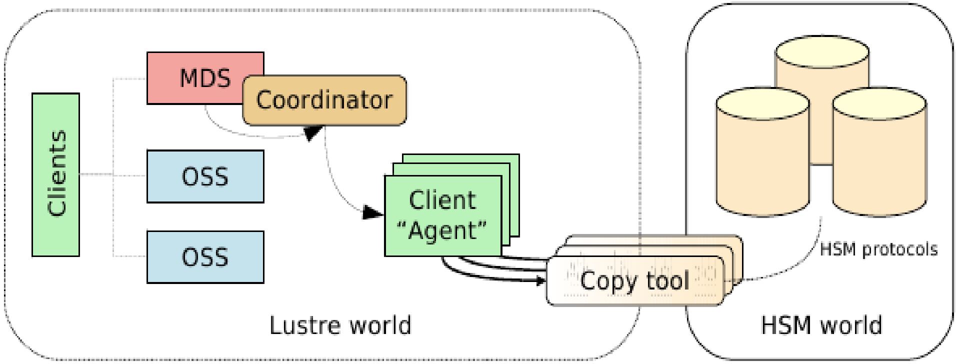
restore

File *my_file* on Lustre

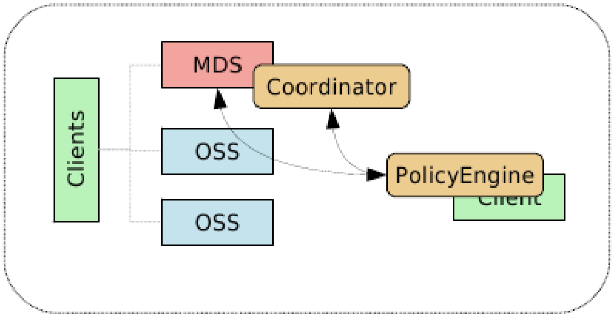
Lustre HSM

- Developed by CEA
- Features
  - Migrate data to an external storage (HSM)
  - Free disk space when needed
  - Bring back data on cache-miss
  - Policy management (migration, purge, soft rm, …)
  - Import from existing backend
  - Disaster recovery (restore Lustre filesystem from backend)
- New components
  - Coordinator
  - Archiving tool (backend specific user-space daemon)
  - Policy Engine (user-space daemon)

Lustre HSM - Architecture



- The coordinator gathers archiving requests and dispatches them to agents
- Agent is a client which runs a copytool which transfers data between Lustre and the HSM



- Policy engine (***robinhood-hsm***) manages pre-migration and purge policies

Lustre HSM - Patches

- Lustre HSM brought by several patches developed by CEA :
  - Adaptation and bugfixes patches (over lustre 2.1.1):
    - LU-810 - Fix helpers for extracting information from HSM changelog records
    - LU-787 – fftruncate blocks when grouplock is done
    - LU - 1072 – Locking bug in grouplock glimpse callback

━ Landed

━ Under inspection

  - Feature patches:
    - Add hsm requests
    - Add hsm flags
    - HSM Posix copytool
    - HSM coordinator
    - Add release feature
    - LU - 827 – Implement a per file data version
    - LU - 941 – Manage dirty flag for hsm-archived files
    - LU - 169 – New layout lock

Lustre HSM – New Layout lock

- LU – 169 - Add a layout lock, a reference counter lsm and a layout generation number

- Patch currently being reworked and split in 4 patches:
  - Layout generation
  - Basic infrastructure for layout lock
  - LSM refcount
  - Core layout lock

- Layout lock opens the doors to:
  - HSM support: releasing and recovering a released file
  - OST rebalancing: move objects between OSTs
  - OST emptying
  - Restriping (Dynamic layouts): allow file layouts to change as the file grows or access patterns change
  - Dynamic layout for subset of a file: restore a part of a file to speed access to critical data
  - Async mirroring: create multiple copies of a file within the same fs namespace

Lustre HSM – Testing

- Currently being tested at:
  - Cines / Prace WP9
  - SGI
  - CEA
  - Bull / HPC R&D labs
  - Bull / IFERC

Lustre HSM – Bull tests

- At Bull's R&D HPC labs, phase 1:
  - Functional tests:
    - Several OSTs: useful testing the new layout lock patch
    - Backend over a local disk (ext4)
    - With robinhood-hsm (policy engine)
  - Helping CEA developers on debugging:
    - Some bugs with the restore functionality and the layout lock
    - Minor bugs in archiving
    - Minor bugs with the archiving tool

Some WA on place but system fully operational now

Lustre HSM – Bull tests

- At Helios (Japan), phase 2:
  - Functional and robustness under high IO load tests:
    - 4 OSS, 60 OSTs
    - High IO load with clients
    - 2 copy agents, backend over storage array
    - 1 robinhood node, mysql db over storage array
  - Debugging:
    - Changelog bugs
    - Statahead issues with Lustre 2.1
    - Copytool load balancing

  No major bugs found but changelog feature needs to be intensively tested

  - Robinhood-hsm tests:
    - Load tests: 3M files
    - Robinhood error recovery
    - Robinhood policies

Lustre HSM – Bull tests

- At Bull's R&D HPC labs, phase 3:
  - Functional, robustness, transition and HA tests:
    - 16 standard clients
    - Multi copy agents configuration over storage array
    - Robinhood HA configuration
  - Validation tests of the Lustre HSM jira tickets
  - Changelog tests
  - Transition tests:

    1) We have a backed-up Lustre filesystem

    2) We want to install Lustre HSM in our already running Lustre filesystem

    3) We do not want to recopy all the data already backed-up

Architect of an Open World™

Lustre HSM – Status & Roadmap

- Lustre HSM compatible client may be supported in Lustre 2.3
- Full Lustre HSM Client (agent and robinhood support) more likely in Lustre 2.4
- Lustre HSM Server targeted to be supported in Lustre 2.4

Lustre HSM – Status & Roadmap – Detail

- Lustre HSM compatible client may be supported in Lustre 2.3

LU - 827 – Implement a per file data version

LU - 941 – Manage dirty flag for hsm-archived files

LU - 169 – New layout lock

- Full Lustre HSM Client (agent and robinhood support) more likely in Lustre 2.4

Add hsm flags

Add hsm requests

HSM Posix copytool

- Lustre HSM Server targeted to be supported in Lustre 2.4

LU - 1333 - Add release feature

HSM coordinator

🟩	Landed in master
🟧	Ongoing work, landing not confirmed
🟥	Patch still to be submitted

Temporary HSM alternative: Lustre backup

- Need to regularly **backup Lustre files**
- Standard Lustre 2.1.1
- No need of releasing files at mid-term (high storage capacity on lustre fs)

The solution: robinhood-**backup**, developed by CEA

Lustre Backup: Robinhood-backup

- Thanks to Lustre changelog, modifications are automatically detected:
  - no need to regularly scan Lustre fs

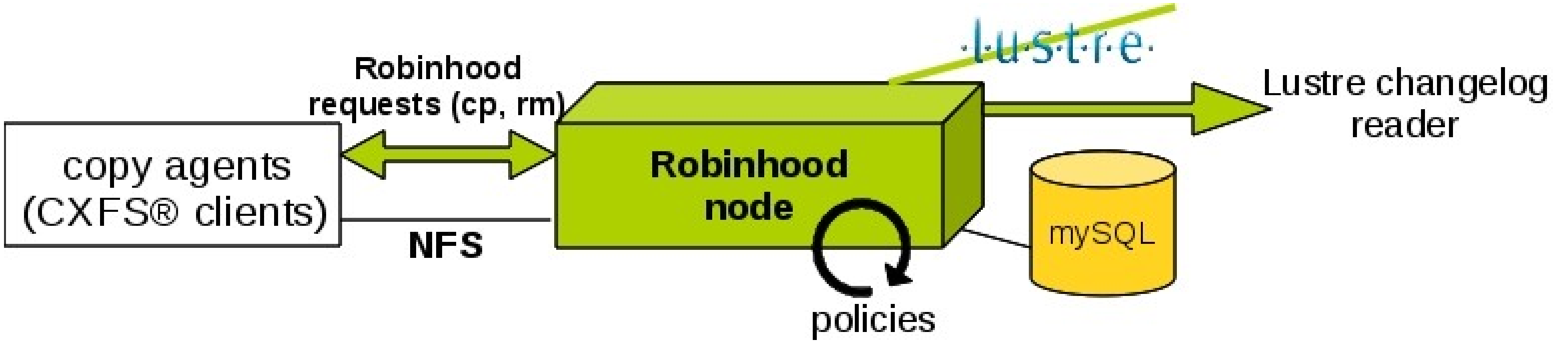
- Policy engine (Robinhood-backup) automatically copying files:
  - Migration policies are defined

- Soft remove on Lustre files
  - Removed files on Lustre are not removed on the backend (also delayed removal)

- Soft transition to Lustre HSM
  - Files are already migrated to the HSM device

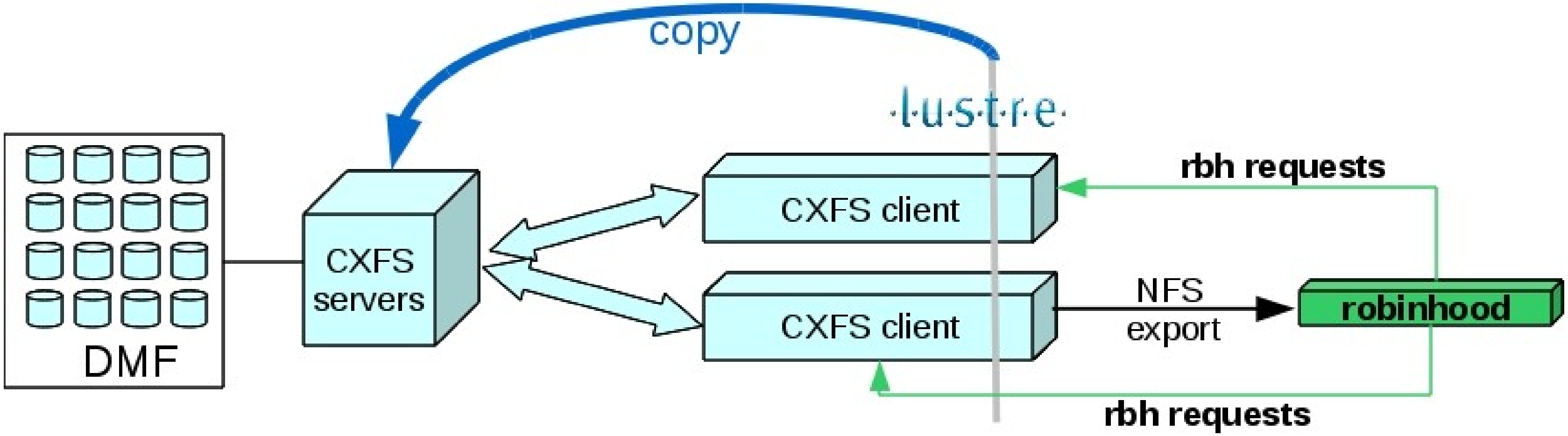
Lustre Backup – Robinhood-backup architecture

- Robinhood-backup:
  - Sees Lustre and DMF contents
  - Registered as changelog reader
  - Manage mySQL database with the state of every file
  - Use wrappers on copy nodes (distant cp and rm commands)

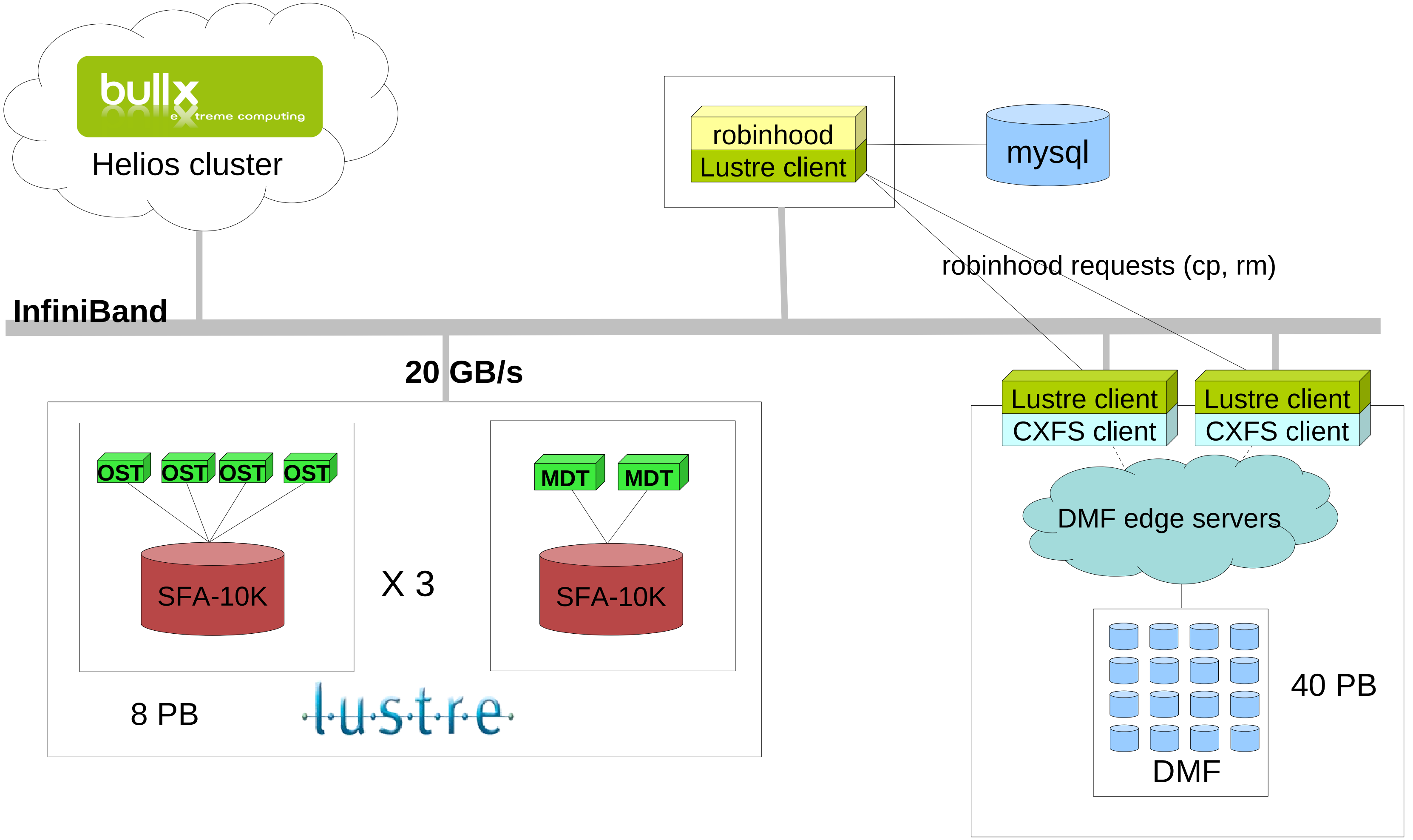


Lustre Backup – Copy agents architecture

- Copy agents:
  - Access to SGI's DMF (CXFS) and Lustre
  - Two or more copy agents (CXFS clients)
  - Robinhood wrapper tool: cp and rm commands sent by robinhood



Lustre Backup - Architecture



Upgrading to Lustre HSM

- What will we have in the future?
  - The need to release some files: from 8 PB to more than 40 PB
  - All the Lustre filesystem already backed-up on DMF
  - Upgrade time limited by the system in production

Backed-up Lustre fs

Soft upgrade

Lustre HSM fs

- Transition based on update of **lustre_mdt_attrs** for every archived lustre file (kind of new LINK feature):
  - Update *flags* (OK), *archive_number* (OK) and *archived_sum* (to be developed)
  - User command allowing to do this (to be developed). Example:

```
lhsmtool_posix --link <lustre_file> <backend_migrated_file>
```

Conclusion

- Lustre HSM is really on the way: landed code + planned landings
- Lustre HSM tests already running on some sites
- The exascale is coming, Hierarchical IO solution with Lustre on top
- Community development model: EOFS & OpenSFS deeply implied
- Want to see Lustre-HSM in action?

See a proof of concept in one of the LUG breaks bullx

instruments for innovation

powered by Bull