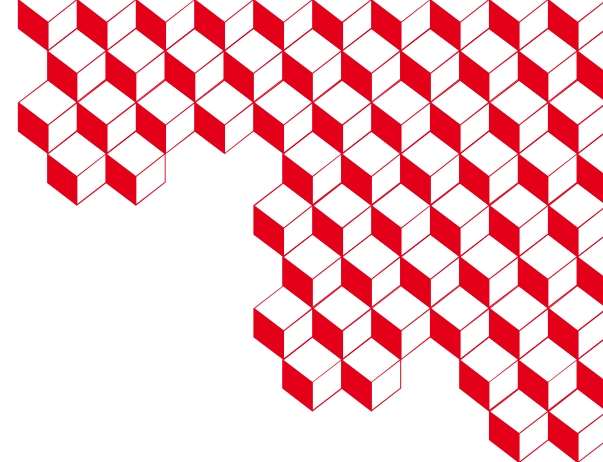




CEA site update

Guillaume Courrier, guillaume.courrier@cea.fr

Sacha Pateyron, sacha.pateyron@cea.fr

Agenda

1. CEA activities
2. HPC compute center at CEA
3. HPC ecosystem : Ocean
4. Lustre at CEA
5. Lustre HSM
6. From 2.12 to 2.15
7. What's next ?



1 ■ CEA Activities

CEA Activities



Defense and security



**Nuclear and
renewable
energies**



**Research and
Development**



**Fundamental
Research**



**Nuclear
weapons**

missions

Design, development,
manufacturing and
maintenance.

Weapons reliability
and safety guarantee
→ Simulation
Program

Nuclear propulsion



Procurement, dismantling of
the old factories

Materials



Conventional defense



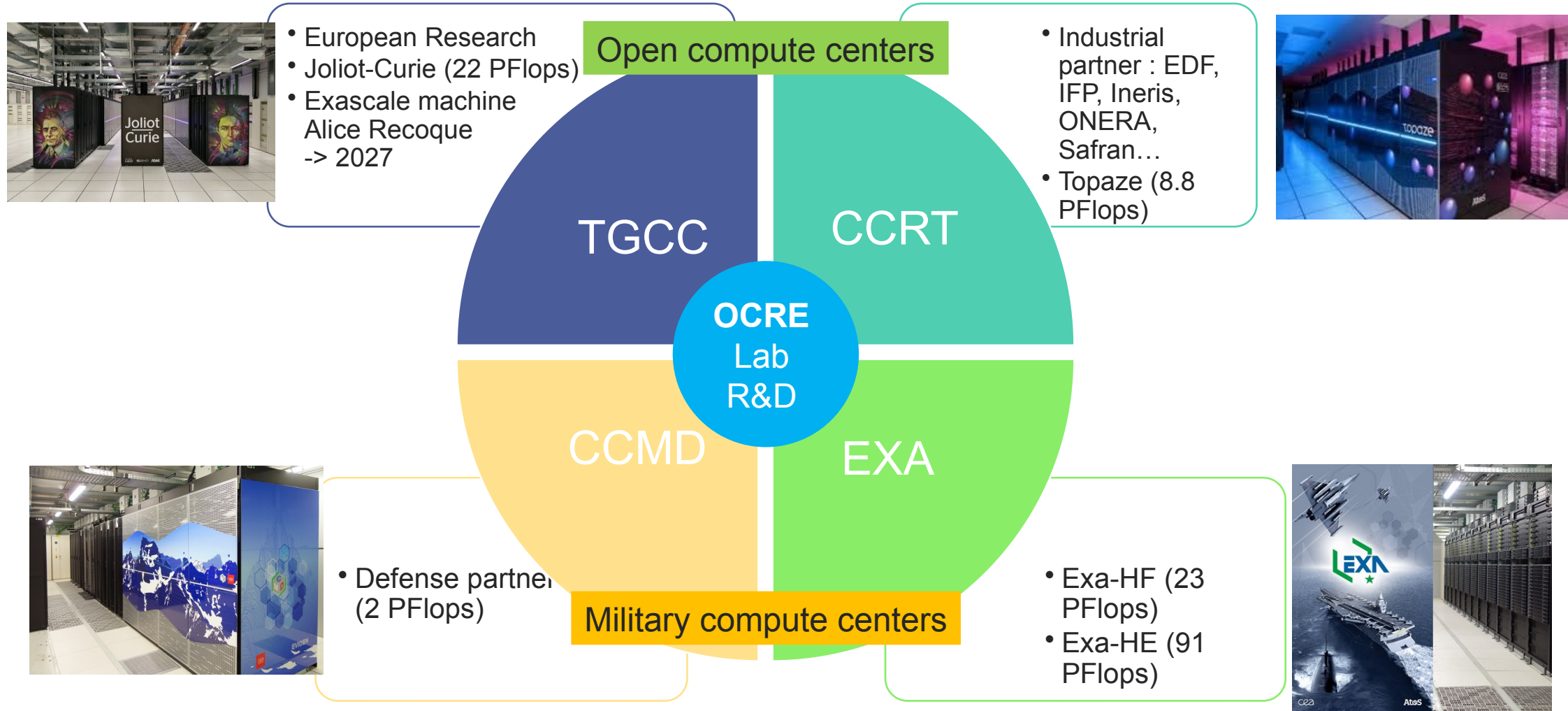
**Security and non-
proliferation**

Non-proliferation, terrorist
risks and threats and
cybersecurity



2 ■ HPC compute centers at CEA

HPC compute centers at CEA

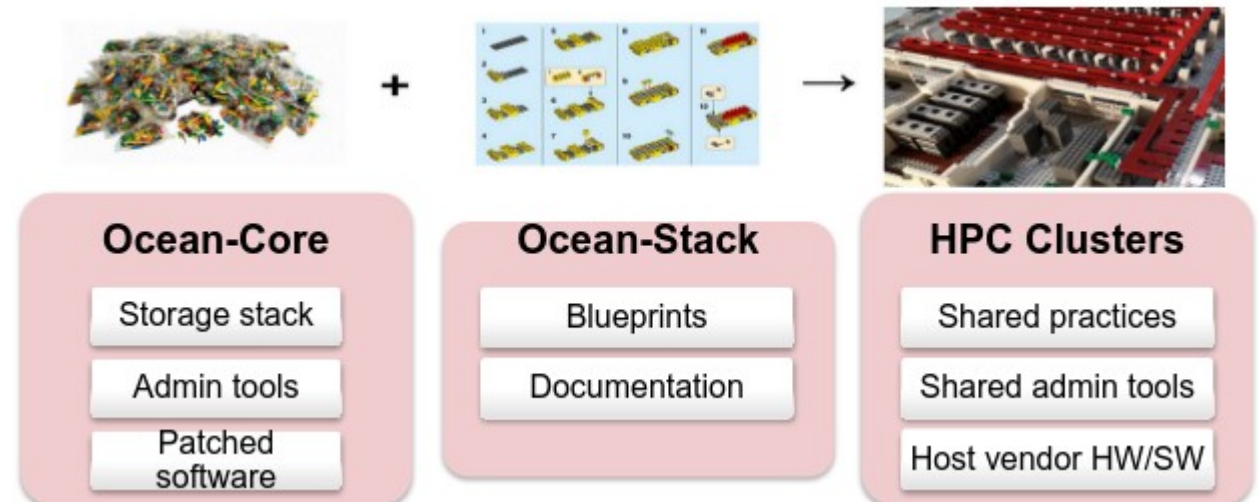




3 ■ HPC ecosystem : OCEAN

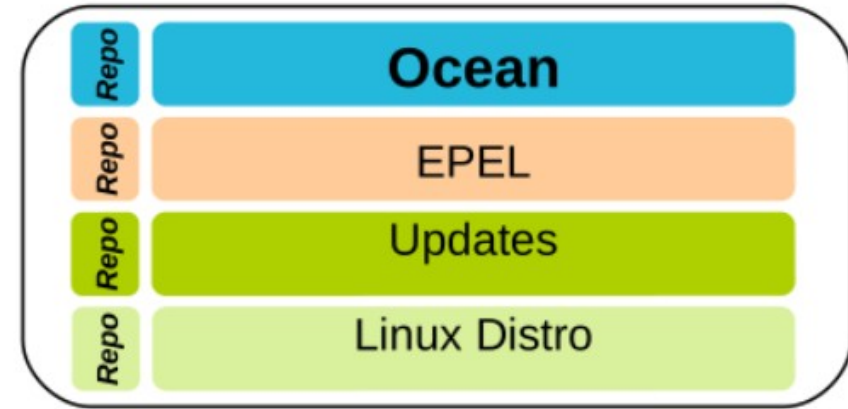
HPC ecosystem : Ocean

- HPC Linux distribution for storage, admin and compute clusters
- Ocean-core / stack :
 - Sharing common administration components
 - Supporting heterogeneous hardware with different software stacks.
- Ocean provides 154 packages :
 - Ocean-stack tools
 - CEA software
 - Up-to-date distribution packages
 - Patched distribution packages
 - Missing dependency and various tools



HPC ecosystem : Ocean

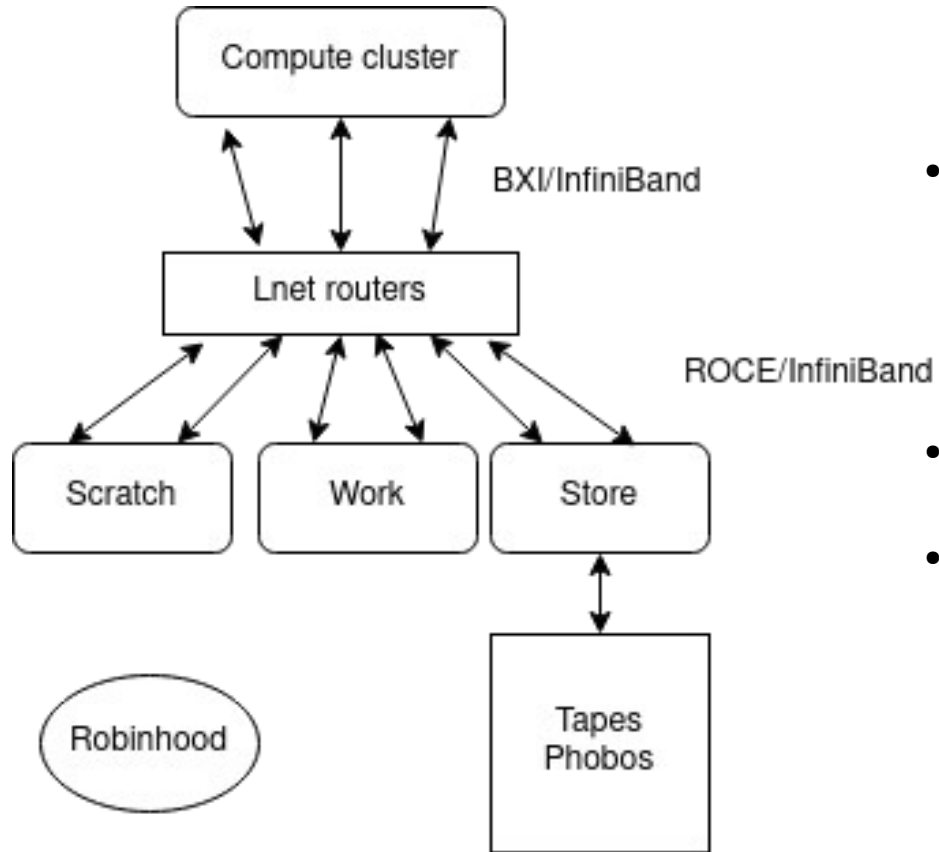
- Based on core Linux distribution
 - 3.X → Almalinux 8.8 / 8.10
 - 4.X → Almalinux 9.4
- Extra repositories → EPEL snapshot + external repository (MOFED)
- Packages are built against this distribution and provided as specific repository → generate an Ocean release
- When Ocean is released, package updates are provided in a dedicated repository (ocean-updates)
- Ocean Forge → Ongoing work to migration CI/CD platform to an open forge → <https://ocean.eupex.eu>





4 ■ Lustre at CEA

Overall architecture



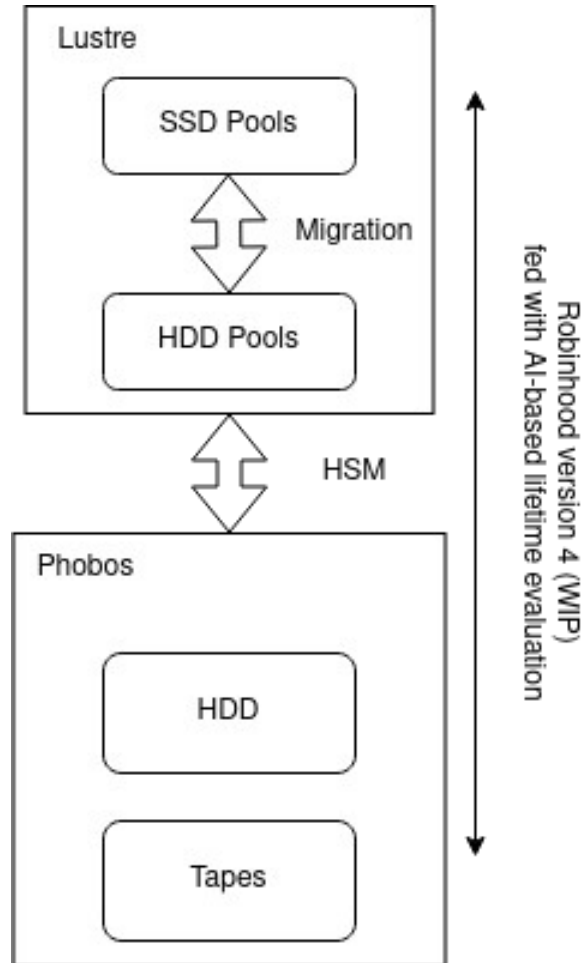
- Infiniband and ROCE Mofed 5.4 / 5.8
CX6 100Gbs → DOCA / 200Gb/s (400Gb/s)
- Scratch : workspace for temporary data
 - designed for throughput and performance
 - purge policy
- Work : permanent workspace (no purge)
- Metrics
 - 168PB (Exa FS + tape)
 - 50M - 400M inodes
 - 1,2TB/s IOR Lustre FS (Exa)
 - 2,8PB weekly production (Exa)

S3 over Lustre

- S3 Object store :
 - Keycloak OpenID authentication → \ ! / warning over specific plugin / need aka. Console UI + Console UI authentication → Aistore product.
 - MinIO architecture → erasure coding → 75 % usable space
 - Disk tier → directory on top of Lustre filesystem.
 - Lustre MDTs using DNEv1 → dedicated MDT



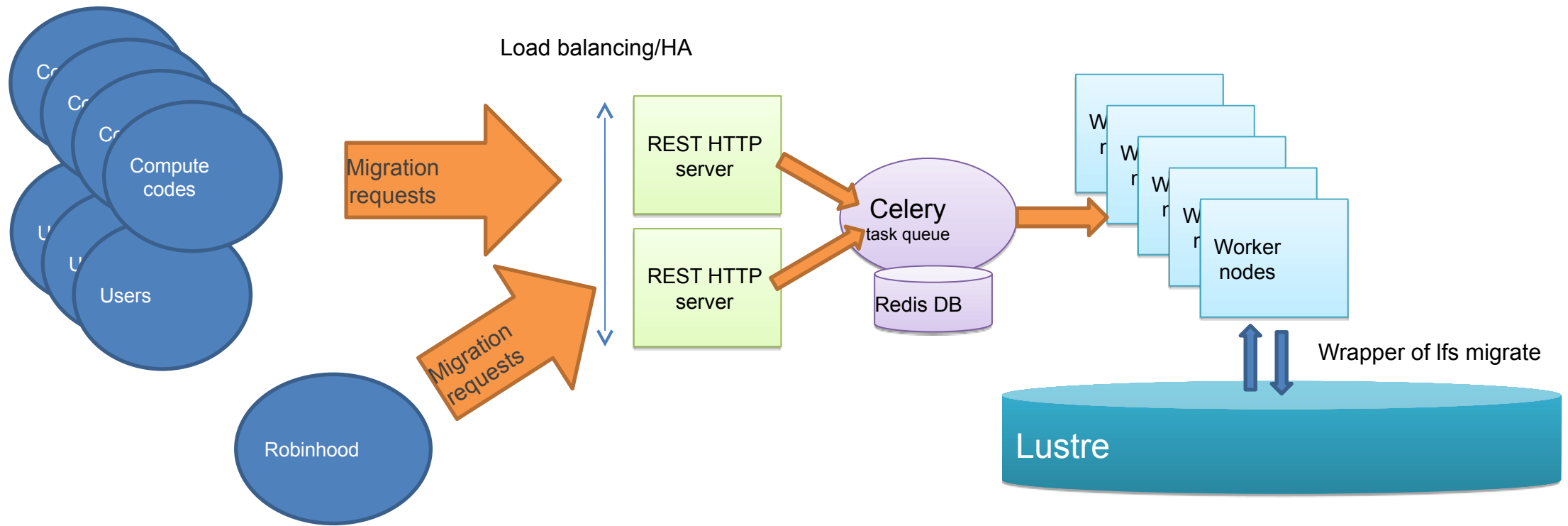
The store filesystem



- Store : long term storage
 - final results
 - designed for data capacity
 - 150+ PB tape storage HPSS and Phobos
 - LTO8 → LTO10
 - Archives from SSD/HDD to HDD on Phobos
 - Later migrated to tapes
- Codes can select the target tier (wrapper of lfs setstripe)
- Default striping to NVMe
- Migration policies from NVMe to HDD and then to tapes (Robinhood)

Pool migrations

- Data migration between NVMe and HDD is performed by a pool of workers





5 ■ Lustre HSM

Lustre HSM

- Manage an external copy of the data to an external storage system

```
$ lfs hsm_archive file  
$ lfs hsm_release file  
$ lfs hsm_restore file  
$ lfs hsm_remove file
```

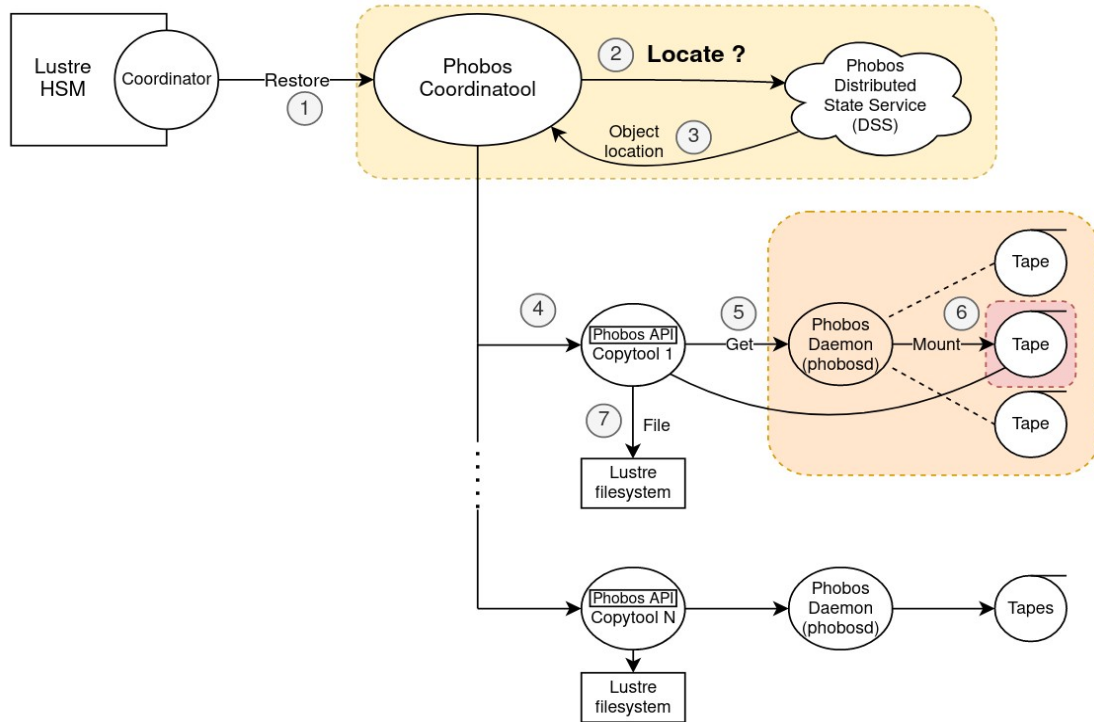
- Automatic data restoration by the client when an application does a read syscall on a released file
- HSM requests are managed by userspace copytools

Long-term storage

- Phobos : Parallel Heterogeneous Object Store
 - Distributed storage system, capable of managing various storage backends (disk, tapes, object stores...)
 - Capable of managing tape libraries (currently moving from HPSS to Phobos)
 - Tape storage based on LTFS (ISO/IEC 20919:2016)
 - Specific optimizations for tape I/Os and mount scheduling
 - Open-source (LGPL v2.1)
 - <https://github.com/phobos-storage/>



Lustre HSM



- Coordinatool
 - <https://github.com/cea-hpc/coordinatool>
 - External coordinator
 - Takes all the HSM requests from the MDT's kernel coordinator
 - Schedules them to the best Phobos node depending on tape position and copytool load
- Phobos
 - <https://github.com/phobos-storage/phobos>
 - <https://github.com/phobos-storage/lustre-hsm-phobos>



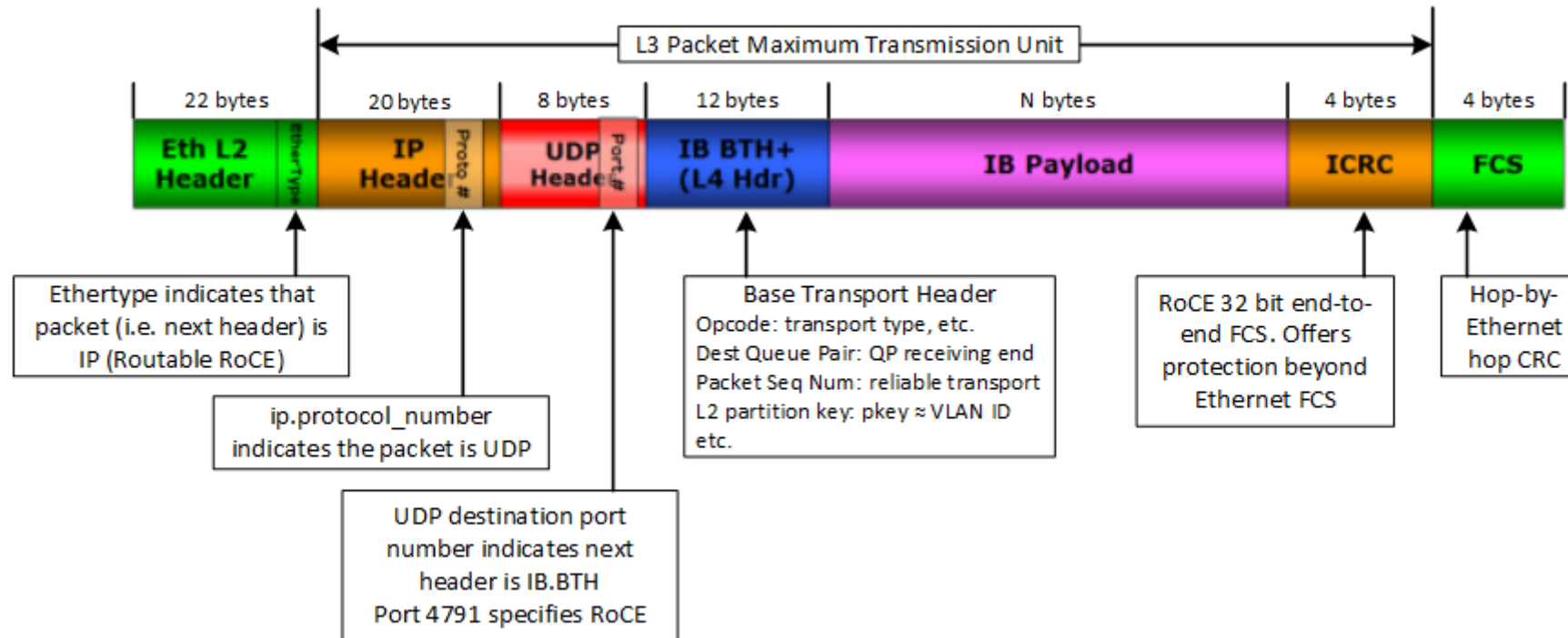
6 ■ From 2.12 to 2.15

From 2.12 to 2.15

- Exa : migrating Lustre 2.12.9.ocean +50PB filesystem to Lustre 2.15.6.ocean
 - File corruption with small read (LU-17482)
 - Memory corruption in FIEMAP ioctl (LU-17013, LU-17110, LU-16480)
 - HSM bugs (LU-17634)
- DNE v3
 - Max inherit to 3 so that each product's content is on the same MDT

From 2.12 to 2.15

- Same time moving from Infiniband to ROCE V2 network :
 - Long network timeout when node is down (LU-17480)
 - Patch in the MOFED drivers to reduce timeout
 - Patch pushed upstream to configure the timeout



From 2.12 to 2.15

- TGCC : hardware EOS → moving from ClusterStore to NVX400X2/3
 - Migrating user data → MPIfileutils :
 - dwalk / dsync / dcmp
 - Initially with no hardlink support
 - Migration of all files with nlinks = 1
 - Use robinhood to get all the FIDs with nlinks > 1
 - Fid2path to get all the paths
 - Final rsync to create missing hardlinks due to trusted.links limitations
 - Softlink where not updated in target if they already exist
 - Remove softlink before migration
 - Both issues fixed now

Lustre in Ocean

- Lustre is packaged in Ocean :
<https://ocean.eupex.eu/download/3.8/ocean-updates/SRPMS/lustre-2.15.6-3.8.ocean3.src.rpm>
 - Old 2.12.9 + 150 patches
 - Current 2.15.6 + 122 patches (some already in 2.15.7)
- All bug fixes and improvements are pushed on the master branch
- Patches are also backported to the 2.15 LTS branch
 - They are only merged into the ocean version once Whamcloud tests succeed



7 ■ What's next ?

What's next ?

- Phobos → Exa / TGCC → full production Q4 2025 / Q1 2026
- Rhel9.4 clients ; Rhel8.10 servers ; DOCA
- Lustre
 - Security features : audit, kerberos and encryption
 - Work on HSM
 - Fix HSM cancel
 - Improve retry and restore
- Renewal of CCRT and Exa supercomputer by 2027
- Copy of a store filesystem
 - Need extra care to manage the HSM state of all the files

What's next ?

- Alice Recoque exaflop supercomputer at TGCC
 - 1 Exaflop/s
 - Data-centric approach
 - +200PB storage system ; +30PB flash
 - + 3TB/s → AI friendly
 - High-Performance Backbone Network → ROCEV2 400Gb/s
 - <https://ec.europa.eu/info/funding-tenders/opportunities/portal/screen/opportunities/tender-details/863893f6-5064-48db-8d96-5e3f30e7ba56-CN>



Thank you

Questions ?

01/10/2025