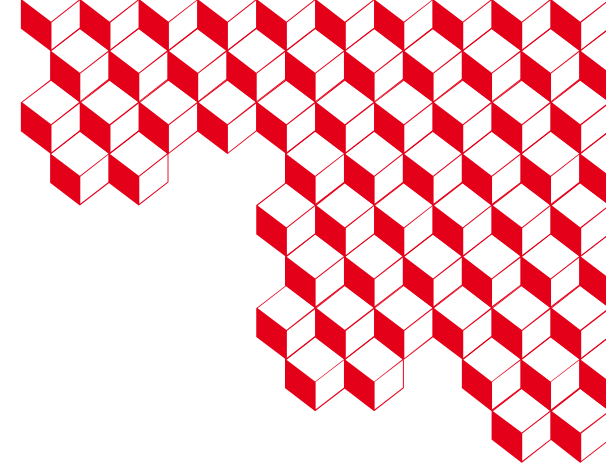




Phobos: an open-source scalable object store optimized for tapes

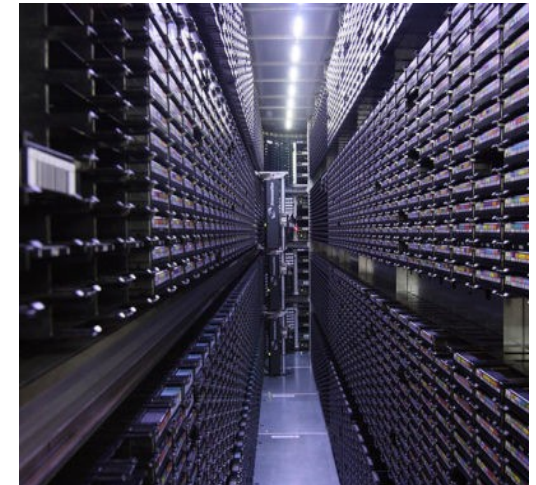
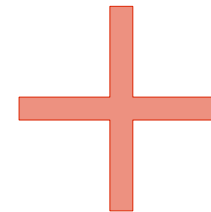
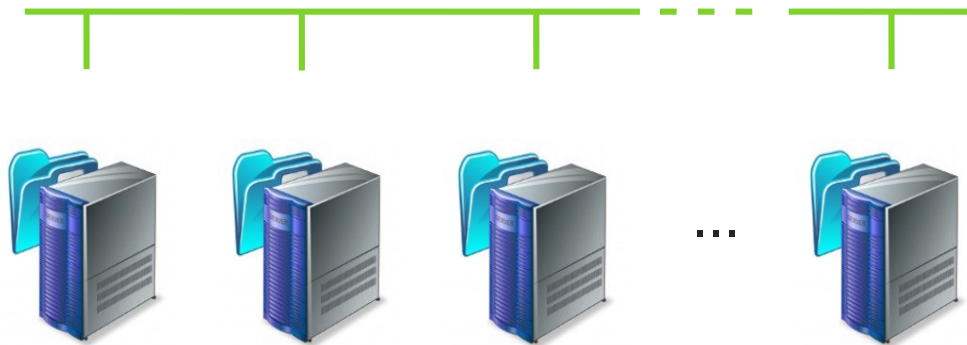
Features overview and Production case at CEA

Rencontres HPSS France 2025

Context

Needs for extremely scalable storage systems at a reasonable price

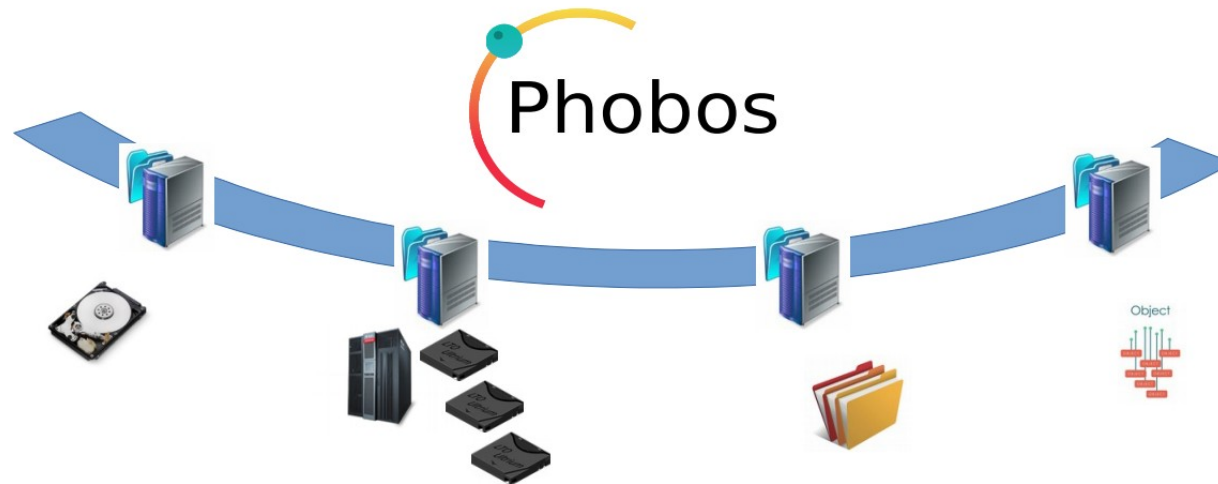
- Object store: horizontal scalability
- Tape library: safe long-term storage at low cost



Phobos: Parallel Heterogeneous Object Store

Goals :

- Manage a distributed set of storage resources on various storage technologies (HDD, tapes, object stores...)
- Implement the best I/O optimizations for each technology without compromise
 - E.g. for tapes: minimize mounts and data sync



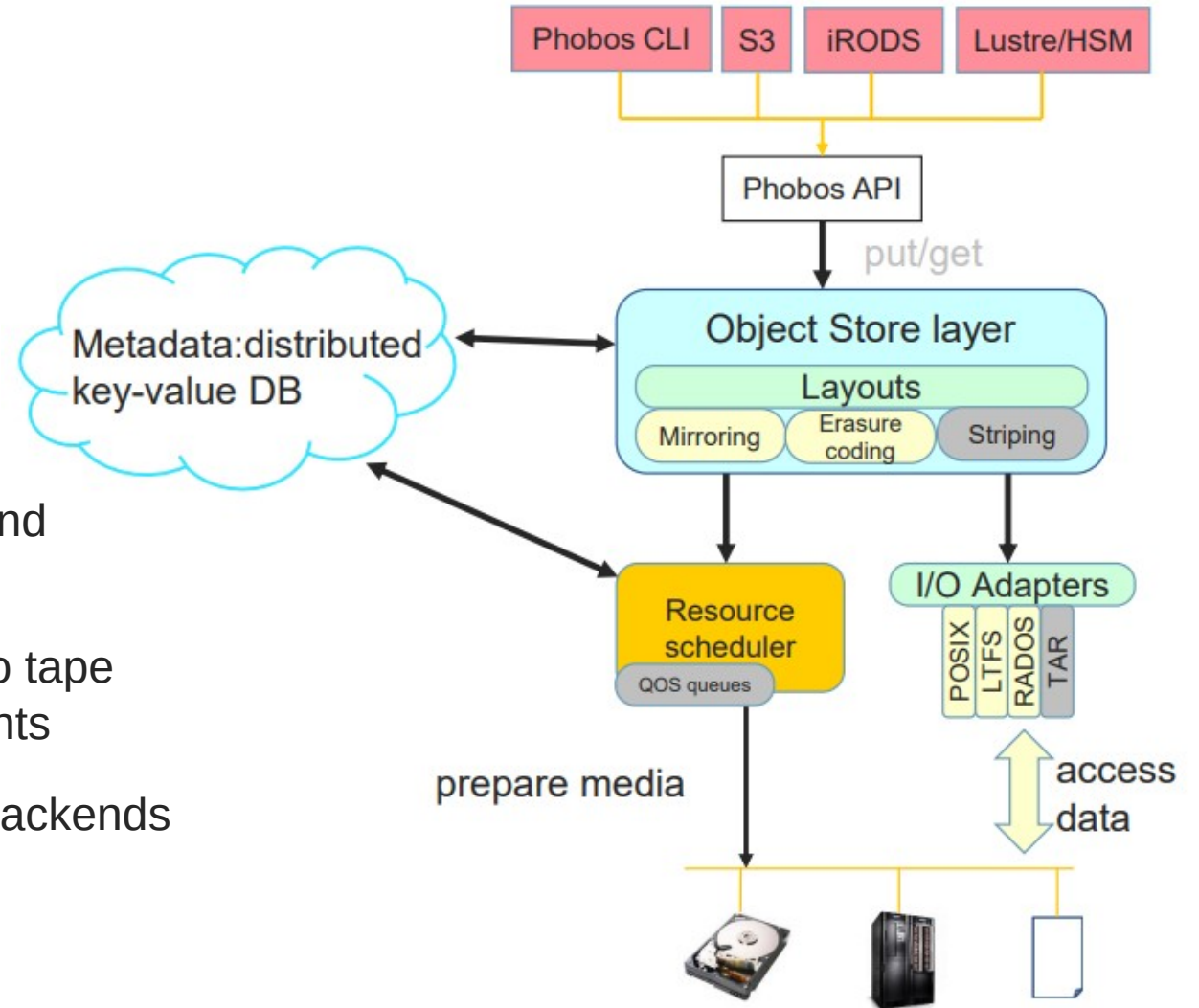
History of the project

- **2014-2015:** development of the initial version
- **2016:** multi-Petabyte of genomics data at TGCC
- **2019:** Phobos made open-source (LGPL v2.1), available on github
- **2022:** Phobos as Lustre/HSM backend at CCMD
- **2023:** First version of Phobos in production outside CEA, Stanford Research Computing, University IT, USA
- **2024-2025:** development of features to deploy Phobos at EXA and TGCC

Phobos: components overview

- **Front-ends:** API
 - Connectors: CLI, S3, iRODS, Lustre/HSM
- **Layout plugins:**
 - RAID1 (mirroring with n copies)
 - RAID4 (2+1 with XOR)

The layout goal is to increase performance and provide fault tolerance
- **Resource scheduler:** optimizes stream to tape drives (grouped read), minimizes tapes mounts
- **I/O adapters:** support of multiple storage backends (POSIX, LTFS, RADOS)





1 ■ Phobos features overview

Phobos features illustrated

Easy setup

- Drive setup

```
phobos drive add --unlock /dev/st1
```

- Tape addition & formatting

```
phobos tape add --type lto9 [073200-073222]L9
```

```
phobos tape format --unlock [073200-073222]L9
```



All done! Phobos is ready for I/Os!

Phobos features illustrated

Adding objects

```
phobos put file/to/put objid
```

Options available

- **Family**

- Directory, tape, ...

```
phobos put --family tape file/to/put objid
```

- **Layouts**

- Raid 1

```
phobos put --layout raid1 --lyt-params repl_count=2 file/to/put objid
```

- Raid 4 (2 + 1)

```
phobos put --layout raid4 file/to/put objid
```

Phobos features illustrated

Object versioning:

- Object uniqueness

```
phobos put /path/to/file objid1
```

→ fails if objid1 exists

- Creating new object version

```
phobos put --overwrite /path/to/file objid1
```

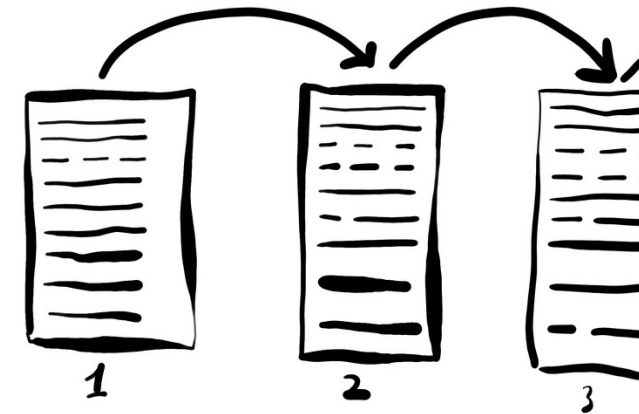
→ creates a new version of objid1

- Listing object versions

```
phobos object list --deprecated objid1
```

- Retrieving an old version

```
phobos get --version 1 objid1 file.out
```

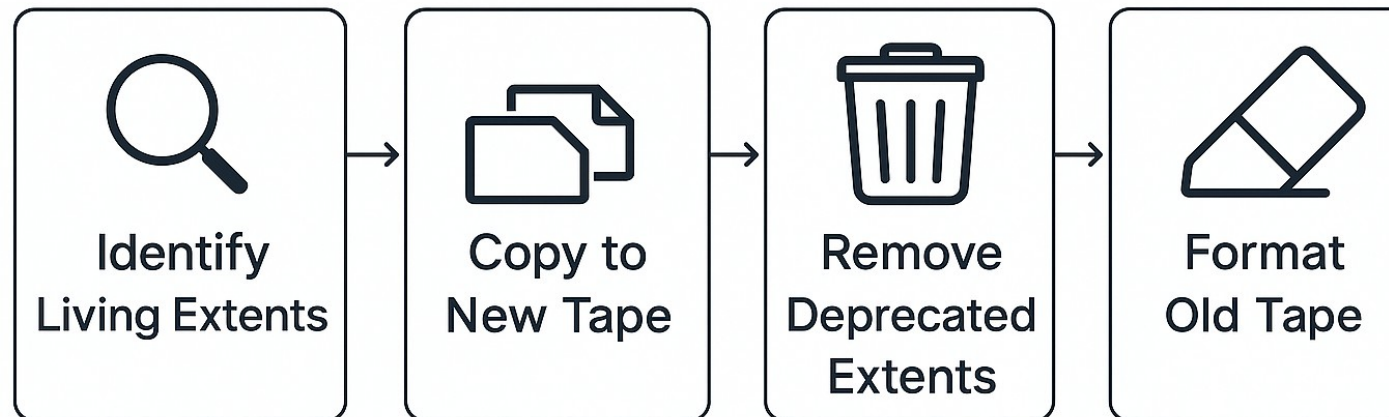


Phobos features illustrated

Repack: tape life cycle

- Optimize tape utilization
- Migrate data between tapes

```
phobos tape repack 073222L9
```



Phobos features illustrated

Resource partitioning with tags:

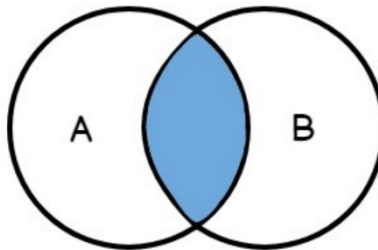
- Tagging resources

```
phobos tape update --tags classA,classB [073000-073099]L9
```

- Pushing data to specific resources

```
# push data to any media with tag "classB"  
phobos put --tags classB /path/to/file objid1
```

```
# push data to a media with both tags "classA" and "classB"  
phobos put --tags classA,classB /path/to/file objid2
```



Phobos IO scheduling

- **Grouped put** : sync après un certain temps, un certain nombre d'objets, ou un certain volume écrit
- **Grouped read** : regroupement des lectures par bande
- **Fair-share dynamique** paramétrable en fonction de la charge par type de requête (format / read / write) : min et max par type de drive et par type de requête
- Historisation des **erreurs SCSI** pour déterminer les composants défectueux (lecteurs, bandes...)



2 ■ Phobos 3.0: The object copy feature

Phobos objects copies

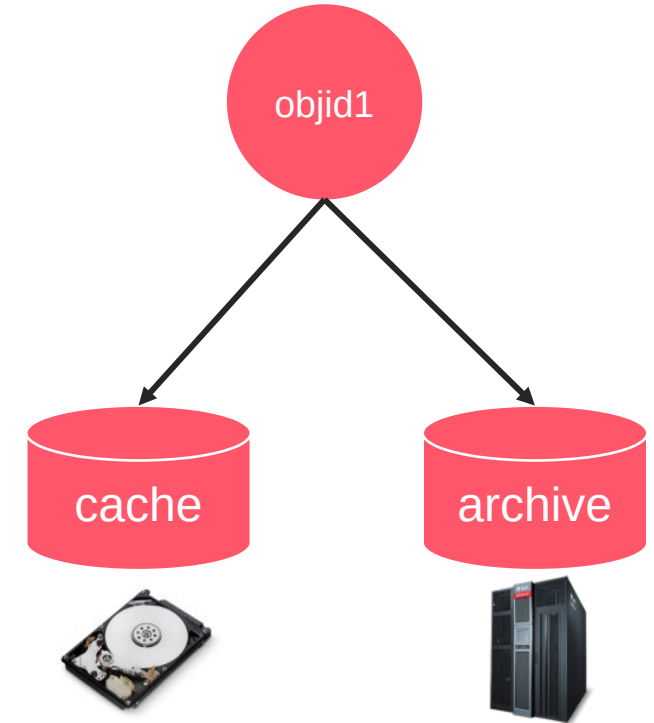
An object can now have multiple copies

This generic feature can be used to

- Migrate data between different storage technologies (E.g. LT08 → LTO10)
- Do internal HSM (Hierarchical Storage Management) policy (dir ↔ tape)
- Do security mirroring between multiple sites

Phobos objects copies

- Copy objects on different medias (POSIX, tape,...)
- A copy of an object represents a copy of its data on a medium (dir, tape, ...).
- Creating a copy doesn't create a new version of an object



Phobos objects copies



- **Create a copy:**

```
phobos copy create objid1 cache
```

→ can take same options as the phobos put (layout, family, ...)

- **Delete a copy:**

```
phobos copy delete objid1 cache
```

- **Listing copies:**

```
phobos copy list objid1
```

Phobos objects copies

A new option **--copy-name** has been added to choose on which copy to work

- Put:

```
phobos put --copy-name cache /path/to/file objid1
```

- Get:

```
phobos get --copy-name cache objid1 file.out
```

The copy name can be linked with put parameters (layouts, family, ...)

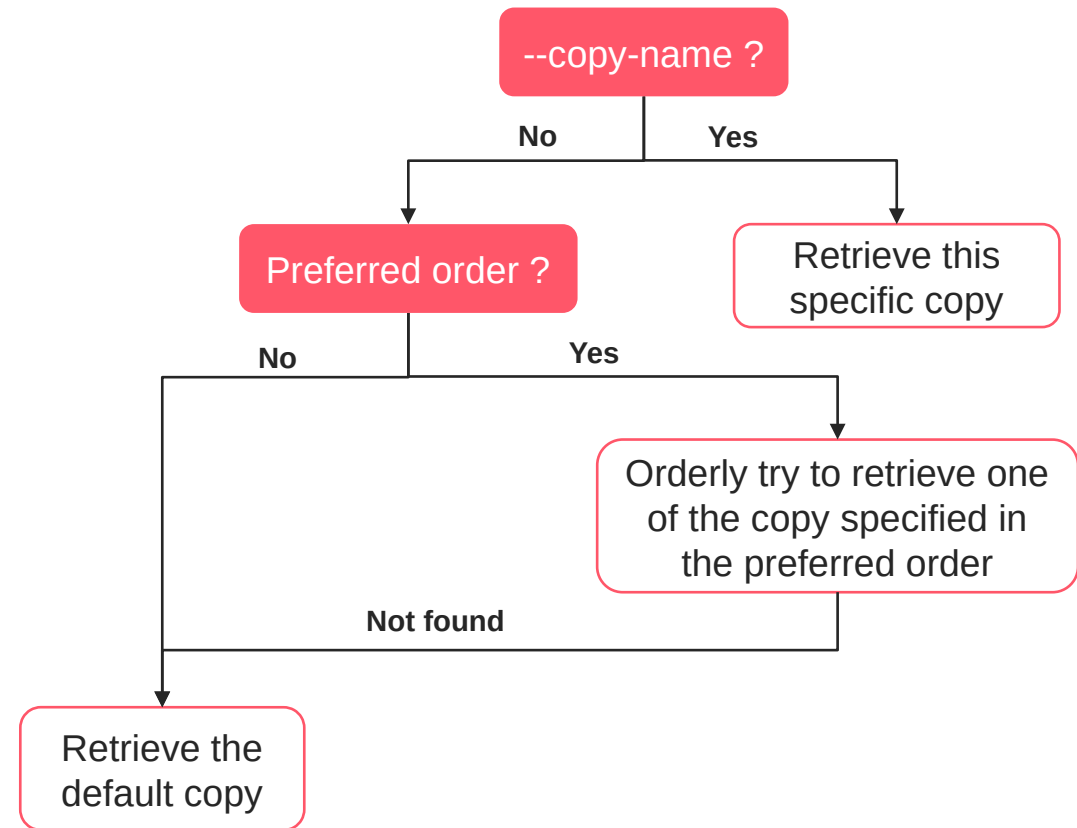
Phobos objects copies

Get preferred order

- A priority ordered list based on copy name when retrieving copies

`get_preferred_order=fast,cache,archive`

- If set, this ordered list will be used first:
get preferred order → default copy
- Can be set in the configuration





3 ■ Production case at CEA

Lustre/HSM – Phobos backend



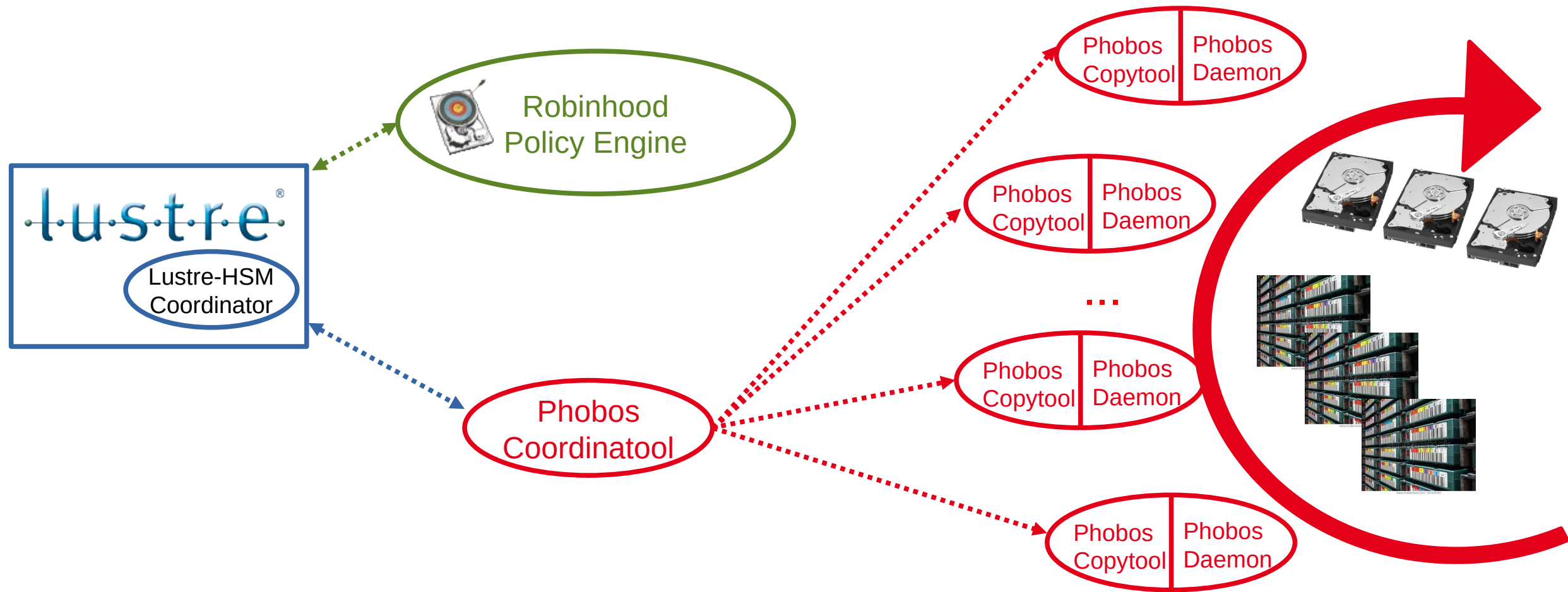
Phobos is deployed as the Lustre/HSM tape storage backend at CEA.

It is integrated into a complete toolchain, including

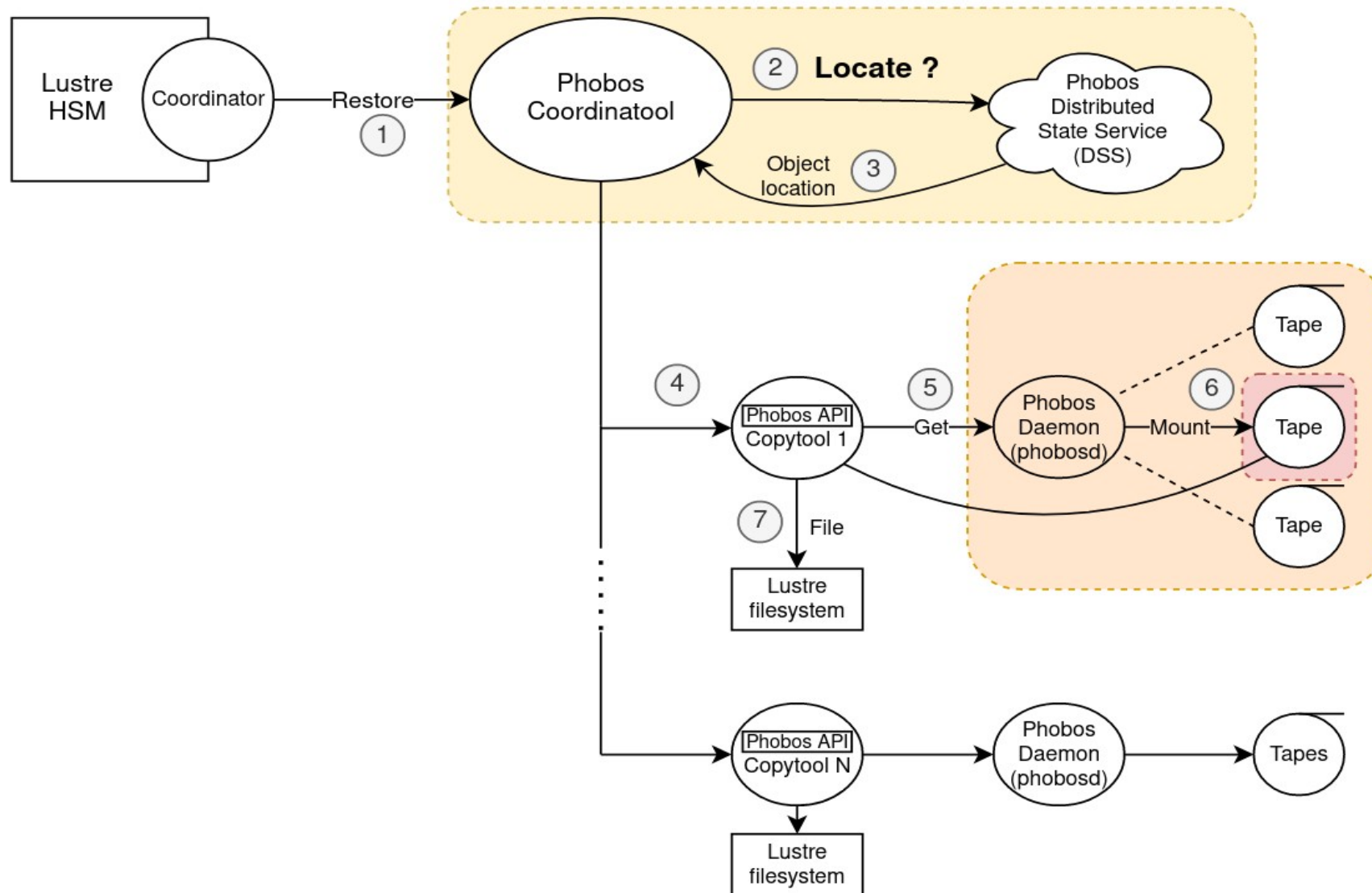
- Phobos Coordinatool: <https://github.com/phobos-storage/coordinatool>
- Phobos Copytool: <https://github.com/phobos-storage/lustre-hsm-phobos>
- Robinhood Policy Engine: <https://github.com/cea-hpc/robinhood>

This toolchains will also be used for the future Lustre/HSM tape backend of new EXA and Alice Recoque supercomputers

Lustre/HSM - Phobos tools chain



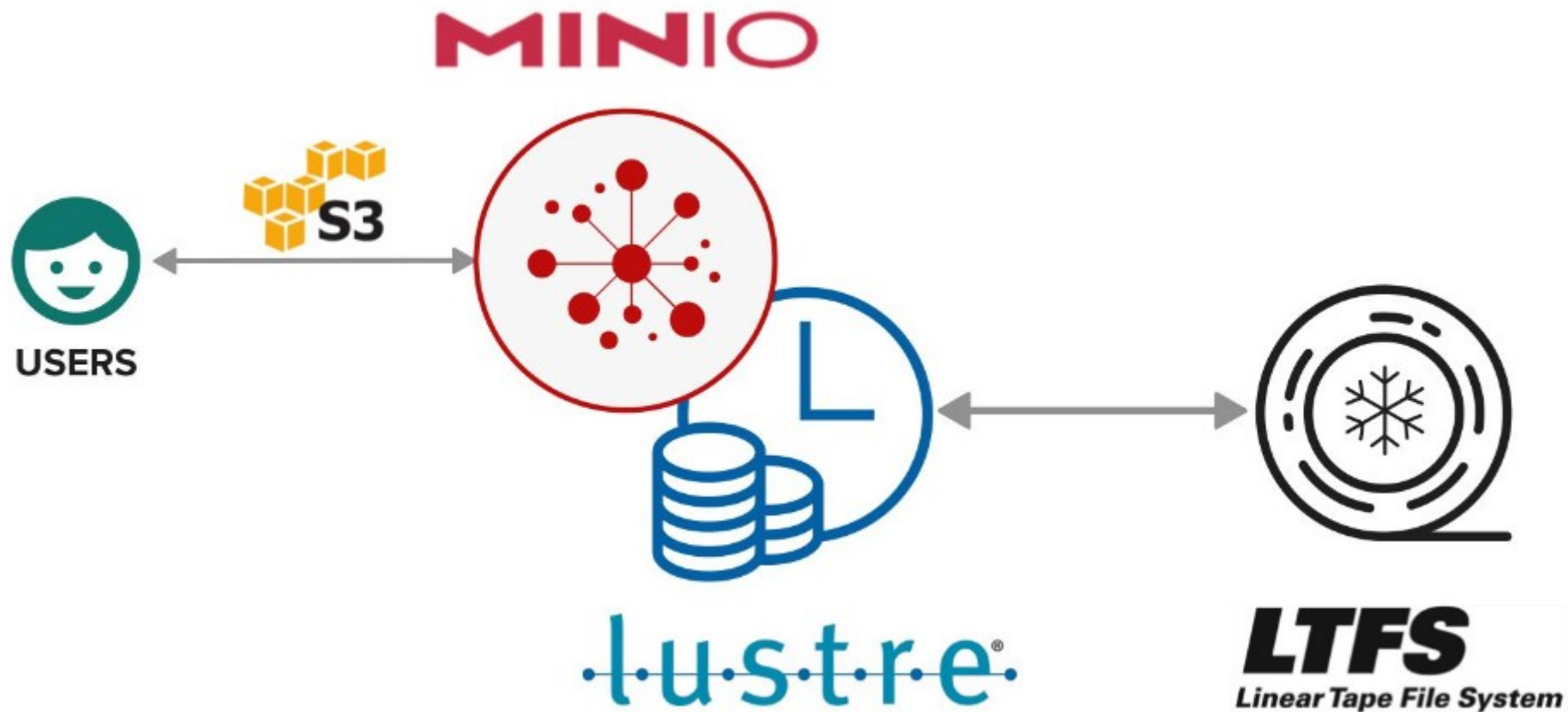
Lustre/HSM - Phobos tools chain



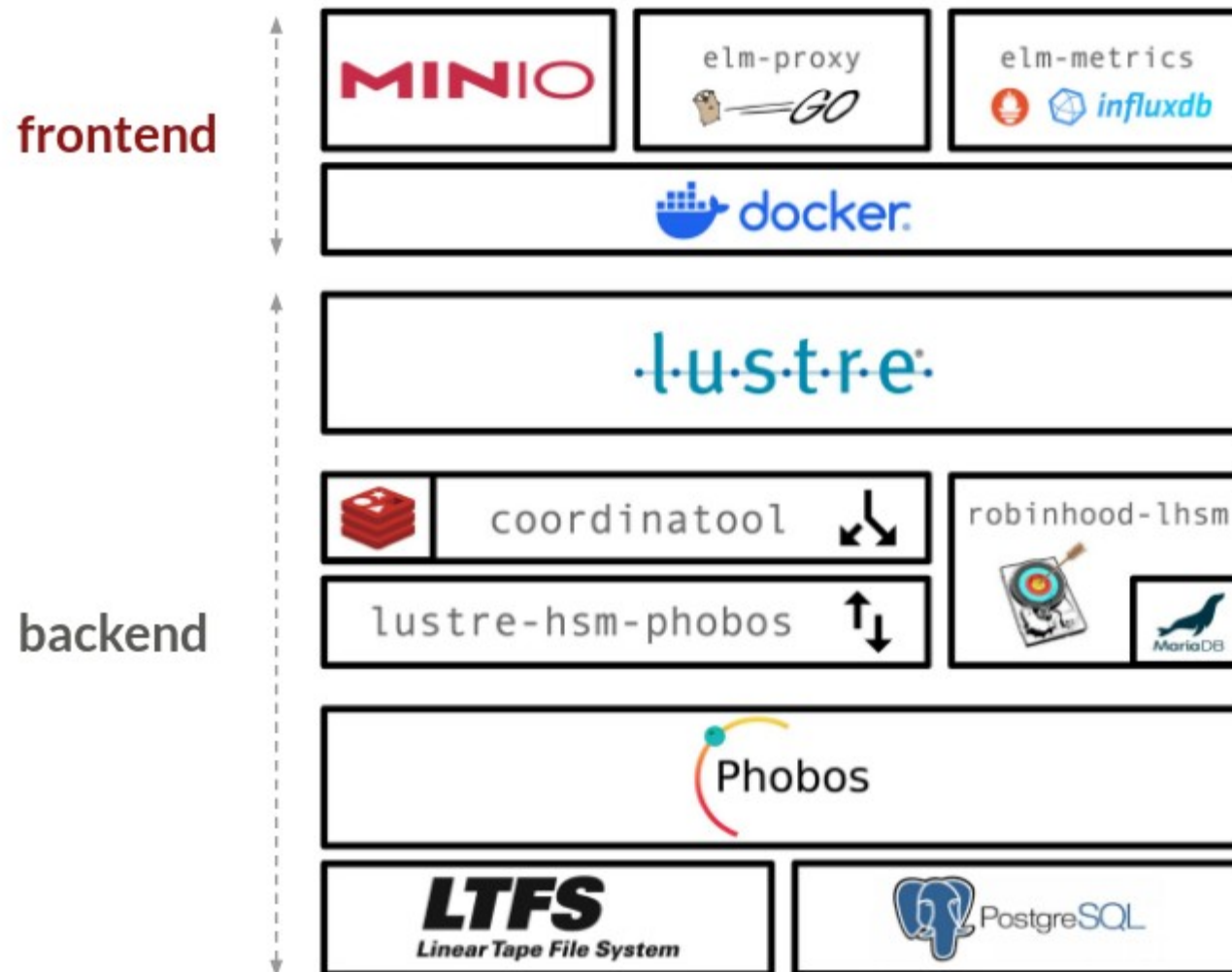


4 ■ Production case at Stanford

Elm@Stanford : big picture



Elm@Stanford : IO stack

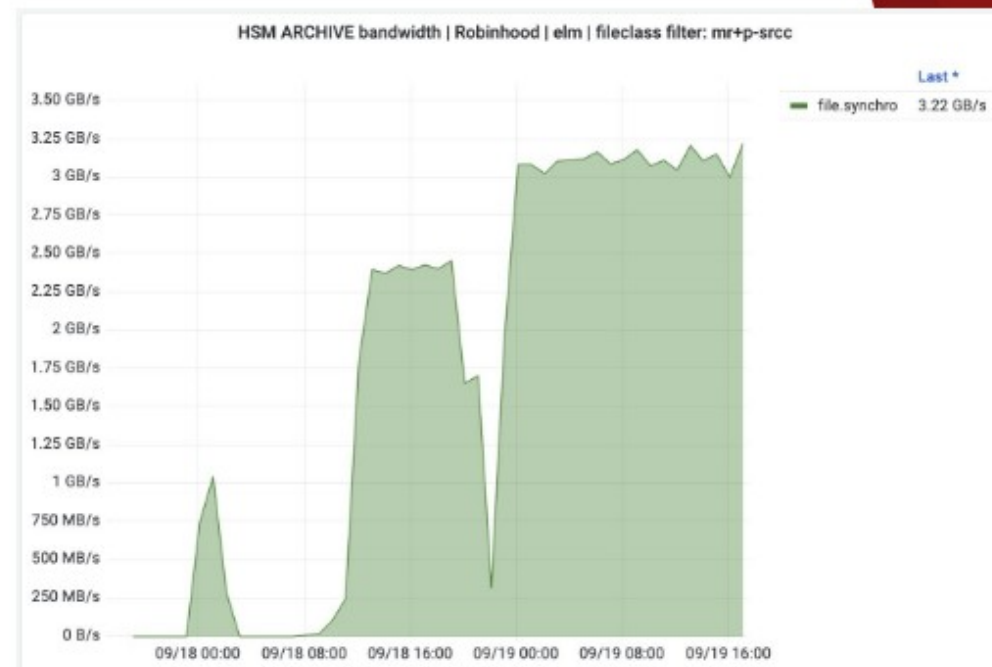
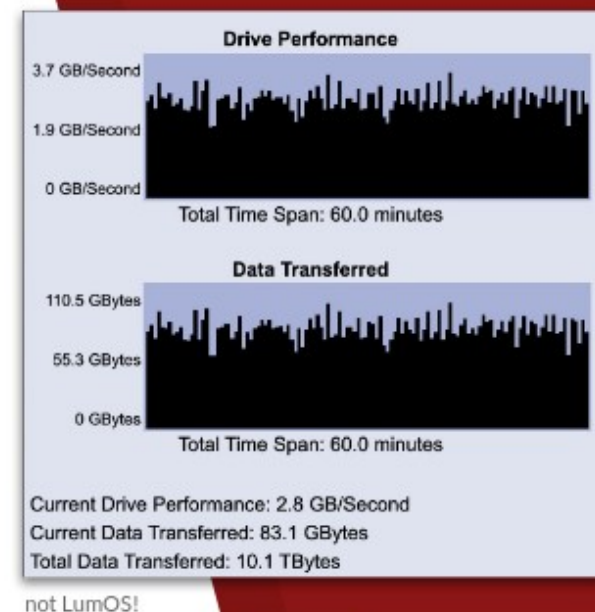


Lustre/Phobos LAD'24 “hero” run

- ▶ Lustre/HSM with DNE archive run using large files, 4 data movers (one per MinIO shard), 4 LTO-9 SAS drives per mover
- ▶ Allow all drives to write:

```
# clush -w @dm phobos sched fair_share --type LT09 --max 0,4,0
```

- ▶ LTFS sync every 2 mins or 1000 files or 16 GiB (set in phobos.conf)
- ▶ hsm/max_requests=750/MDT
- ▶ Results:
 - ▷ 3.22 GB/s aggregate max
 - ▷ 201.25 MB/s per drive
 - ▷ 10+ TB/hour archived
 - ▷ 1 PB in 4 days archived



Elm@Stanford : CLI overview

```
[root@elm-rcf-hn01 ~]# clush -w@dm -x elm-ent-dm05 -b phobos drive status
```

elm-ent-dm01

address	currently_dedicated_to	device	media	mount_path	name	ongoing_io	serial
0	W	/dev/sg2	010050L9	/mnt/phobos-sg2	/dev/tape/by-id/scsi-10110057FB	True	10110057FB
1	W	/dev/sg12	010052L9	/mnt/phobos-sg12	/dev/tape/by-id/scsi-10120057FB	True	10120057FB
2	W	/dev/sg1	010023L9	/mnt/phobos-sg1	/dev/tape/by-id/scsi-10130057FB	True	10130057FB
3	W	/dev/sg11	010041L9	/mnt/phobos-sg11	/dev/tape/by-id/scsi-10140057FB	True	10140057FB

elm-ent-dm02

address	currently_dedicated_to	device	media	mount_path	name	ongoing_io	serial
4	W	/dev/sg4	010045L9	/mnt/phobos-sg4	/dev/tape/by-id/scsi-10210057FB	True	10210057FB
5	W	/dev/sg13	010013L9	/mnt/phobos-sg13	/dev/tape/by-id/scsi-10220057FB	True	10220057FB
6	W	/dev/sg5	010019L9	/mnt/phobos-sg5	/dev/tape/by-id/scsi-10230057FB	True	10230057FB
7	W	/dev/sg15	010012L9	/mnt/phobos-sg15	/dev/tape/by-id/scsi-10240057FB	False	10240057FB

elm-ent-dm03

address	currently_dedicated_to	device	media	mount_path	name	ongoing_io	serial
8	W	/dev/sg7	010021L9	/mnt/phobos-sg7	/dev/tape/by-id/scsi-10310057FB	True	10310057FB
9	W	/dev/sg17	010068L9	/mnt/phobos-sg17	/dev/tape/by-id/scsi-10320057FB	True	10320057FB
10	W	/dev/sg6	010044L9	/mnt/phobos-sg6	/dev/tape/by-id/scsi-10330057FB	True	10330057FB
11	W	/dev/sg16	010069L9	/mnt/phobos-sg16	/dev/tape/by-id/scsi-10340057FB	True	10340057FB

elm-ent-dm04

address	currently_dedicated_to	device	media	mount_path	name	ongoing_io	serial
12	W	/dev/sg7	010042L9	/mnt/phobos-sg7	/dev/tape/by-id/scsi-10410057FB	True	10410057FB
13	W	/dev/sg17	010037L9	/mnt/phobos-sg17	/dev/tape/by-id/scsi-10420057FB	True	10420057FB
14	W	/dev/sg8	010038L9	/mnt/phobos-sg8	/dev/tape/by-id/scsi-10430057FB	True	10430057FB
15	W	/dev/sg18	010065L9	/mnt/phobos-sg18	/dev/tape/by-id/scsi-10440057FB	True	10440057FB

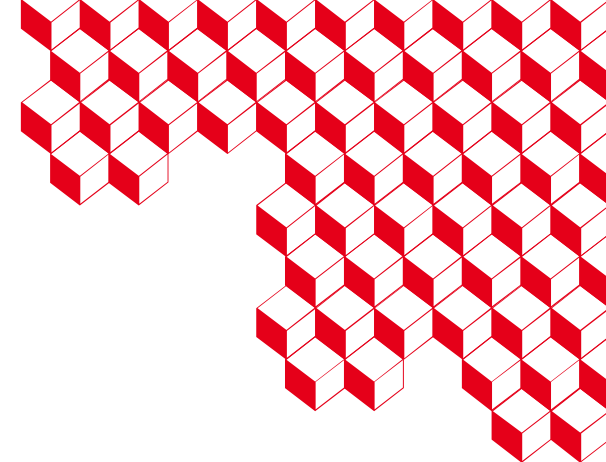


Contribution

Interested?

- Start here: <https://github.com/phobos-storage>
- Documentation and tutorials are available in Phobos's wiki
- Contributions are welcome, as well as feedback!





Thank you !

Gauthier EVRAERD, gauthier.evraerd@cea.fr

Sébastien GOUGEAUD, sebastien.gougeaud@cea.fr

Patrice LUCAS, patrice.lucas@cea.fr

Yoann VALERI, yoann.valeri@cea.fr