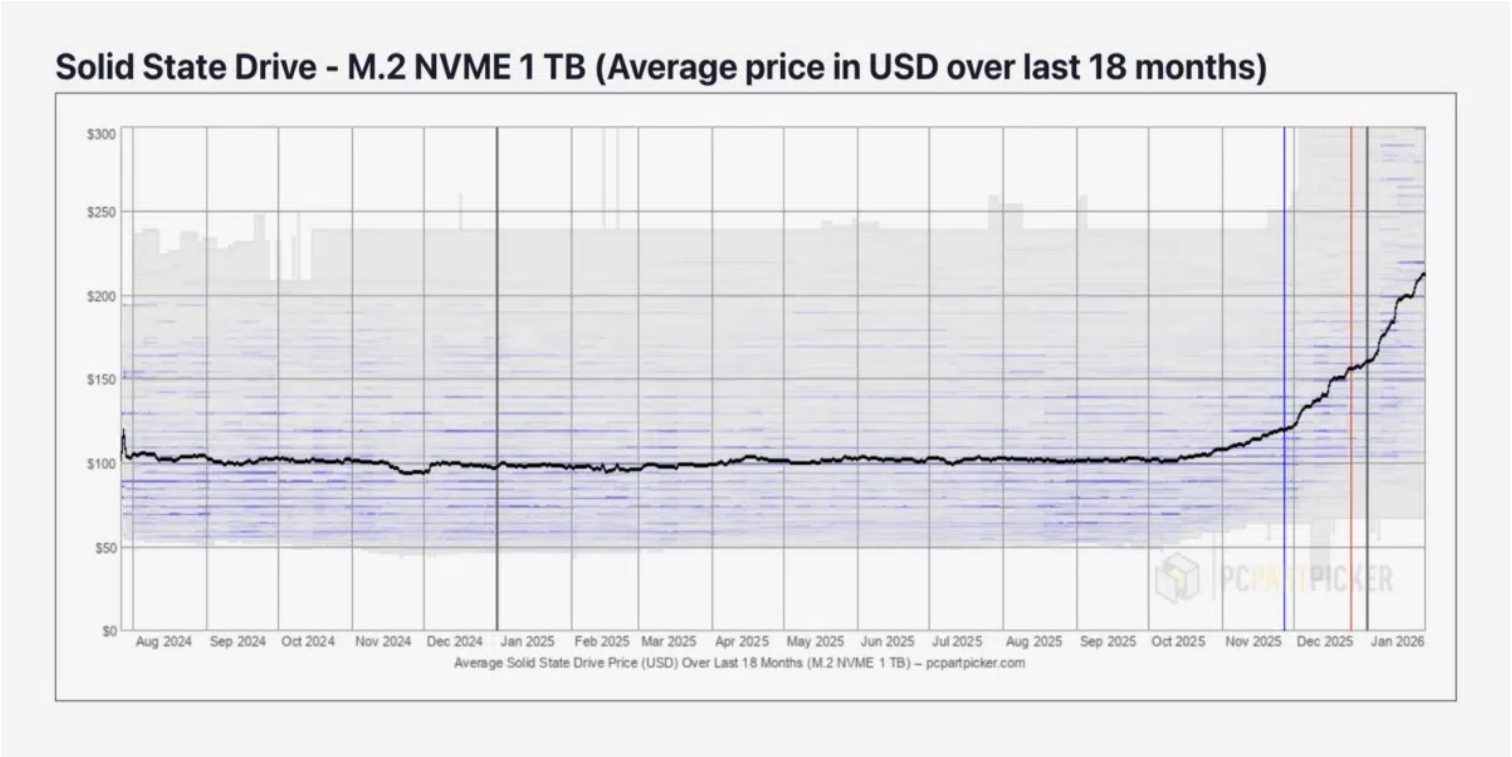
A flexible open-source tape storage system for HPC and more

Thomas LEIBOVICI, CEA thomas.leibovici@cea.fr

CSM-FR 2026, Villeurbanne, France – 9-11 juin 2026

Context



product	September 15, 2025	October 15, 2025	November 14, 2025	December 14, 2025	January 14, 2026	Difference (September to present)
Seagate IronWolf NAS HDD + Rescue 4TB	93.94	97.89	102.94	117.80	131.90	40.41%
Toshiba Cloud Scale Capacity MG10F AFA 22TB	336.47	358.96	425.0	473.99	558.99	66.13%
Western Digital WD Red Plus 8TB	169.9	186.9	198.99	203.99	255.00	50.09%
Toshiba Cloud Scale Capacity MG10ACA 20TB	309.59	329.9	387.99	434.90	479.90	55.01%
Seagate Exos 28TB (recertified)	384.27	366.28	599.0	no offer	559.00	45.47%
Western Digital WD Red Plus 4TB	99.89	104.71	115.89	114.99	133.90	34.05%
Seagate IronWolf NAS HDD + Rescue 8TB	155.9	176.45	193.66	216.00	234.90	50.67%
Seagate BarraCuda 24TB	307.9	369.15	399.0	533.08	499.90	62.36%
Seagate Exos M 30TB	589.07	584.66	612.58	697.90	755.90	28.32%
Toshiba Cloud Scale Capacity MG11ACA 24TB	386.09	442.99	469.8	573.73	602.00	55.92%
Toshiba Cloud Scale Capacity MG09ACA 18TB	287.99	311.98	335.0	366.29	418.99	45.49%
Seagate IronWolf Pro NAS HDD + Rescue 16TB	312.99	340	381.74	384.99	384.99	23.00%
					Average price increase	46.41%

Context

HPC and AI require :

- Extreme performance (TBs/sec, millions of IOPS...)
- High capacities for long term storage (1EB becomes common)
- ...to fit within the budget!

- Tiered storage architecture is a good solution to meet these requirements
- Parallel file systems for the I/O performance
- Without an unlimited budget, tape storage is unavoidable
  but not managed by parallel file systems

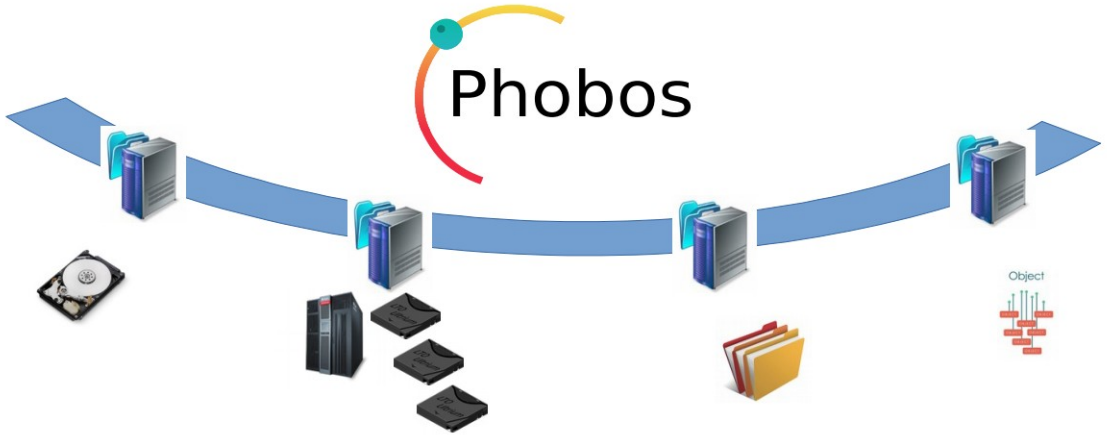
➔ Need for a scalable long-term storage for HPC that makes it easy to integrate and manage tapes

~2€/TB

Phobos: Parallel Heterogeneous Object Store

Goals :

- **Versatility:** Manage a distributed set of storage resources on various storage technologies (disks, tapes, object stores...)
- **Performance**: Implement the best I/O optimizations for each technology
  - Especially for tapes: minimize tape mounts and data sync
- **Sovereignty**: Open formats, no license to access data, no vendor lock-in



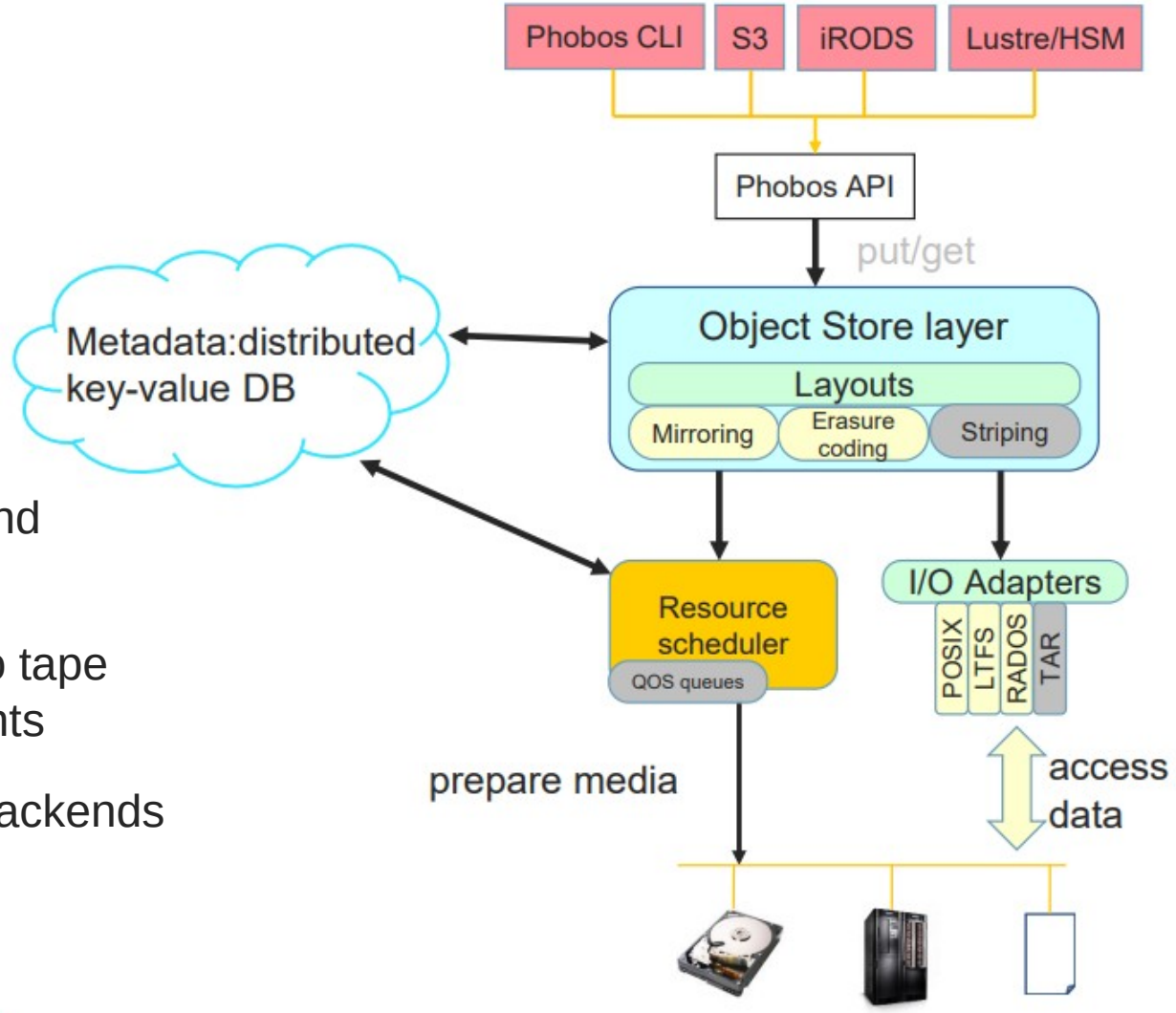
History of the project

- **2014-2015:** development of the initial version
- **2016:** multi-Petabyte of genomics data at CEA/TGCC
- **2019:** Phobos made open-source (LGPL v2.1), available on github
- **2022:** Phobos as Lustre/HSM backend on a medium-scale compute center at CEA
- **2023:** First version of Phobos in production outside CEA, Stanford Research Computing, USA
- **2026:** Migration to Phobos on large scale compute centers at CEA (EXA, TGCC, CCRT, Alice Recoque)

Phobos internal architecture

- **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)



Phobos processes

Core components :

- **phobosd**: run on all I/O nodes
- **TLC** (Tape Library Controller): Single process (possibly with HA) to pilot a library

Front-end processes:

- S3 servers
- lhsmd_phobos: Lustre/HSM copytools
  + coordinatool
- IRoDS connectors
- ...

1. Production case at CEA

Lustre/HSM – Phobos backend

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

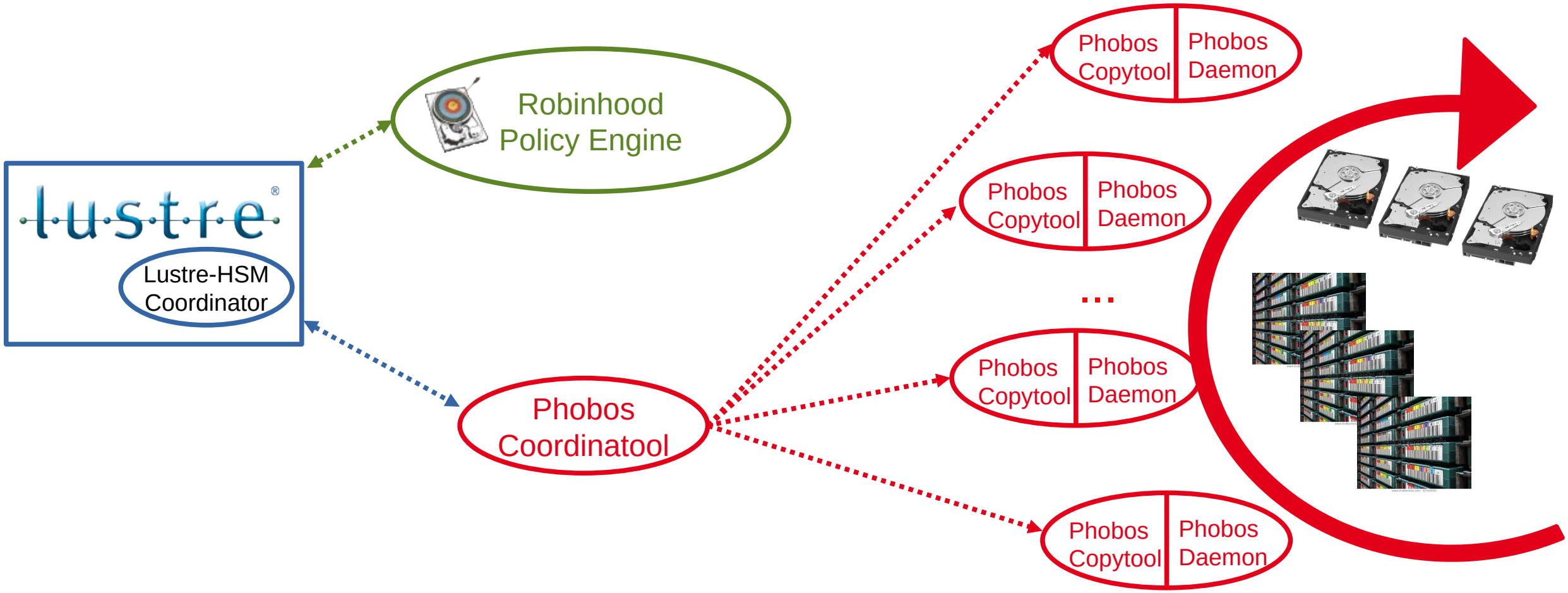
It is integrated into a complete toolchain, including

- 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

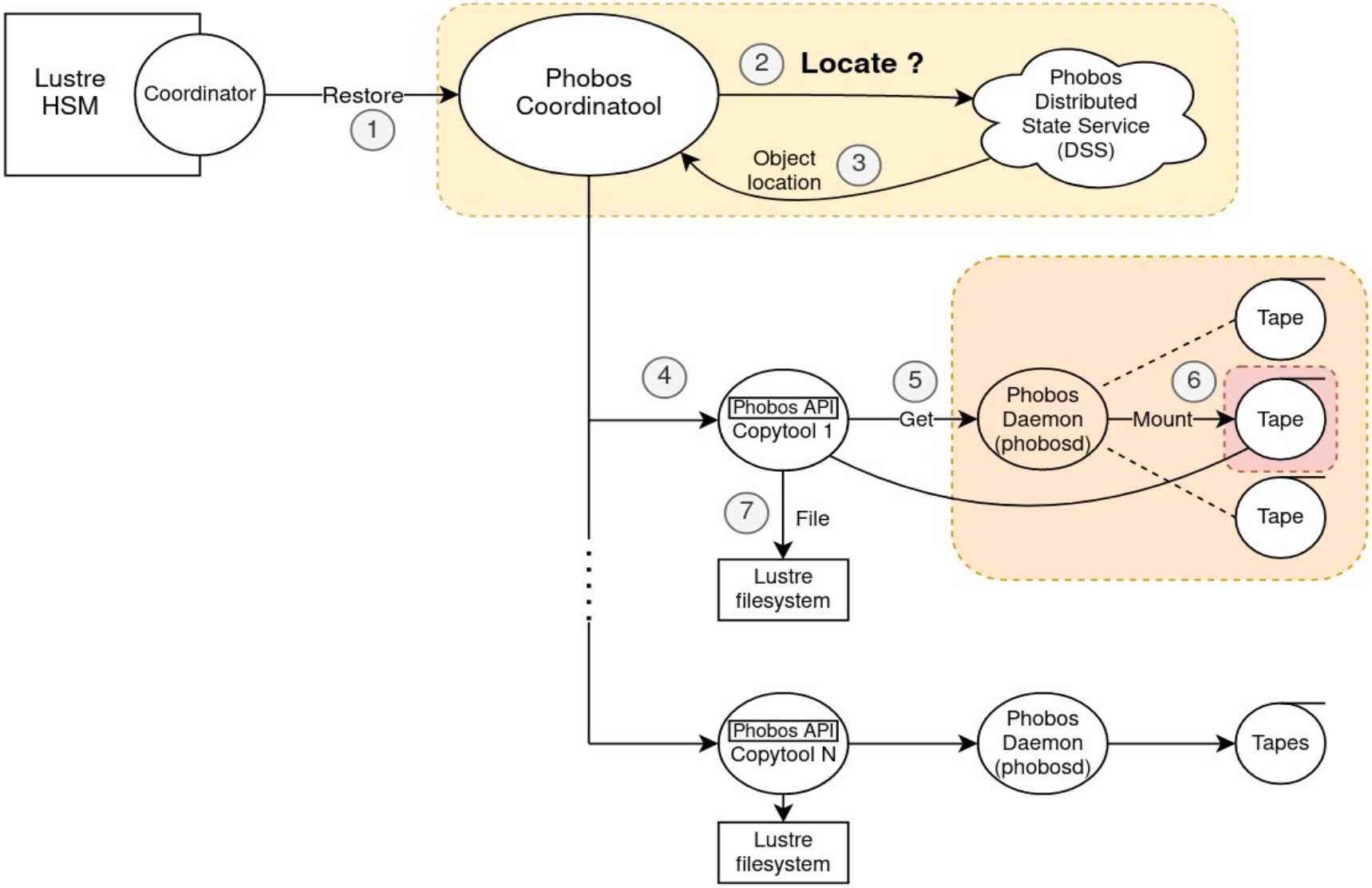
Currently deployed at CCMD and EXA

Soon to be deployed on TGCC, Alice Recoque and CCRT

Lustre/HSM – Phobos tools chain

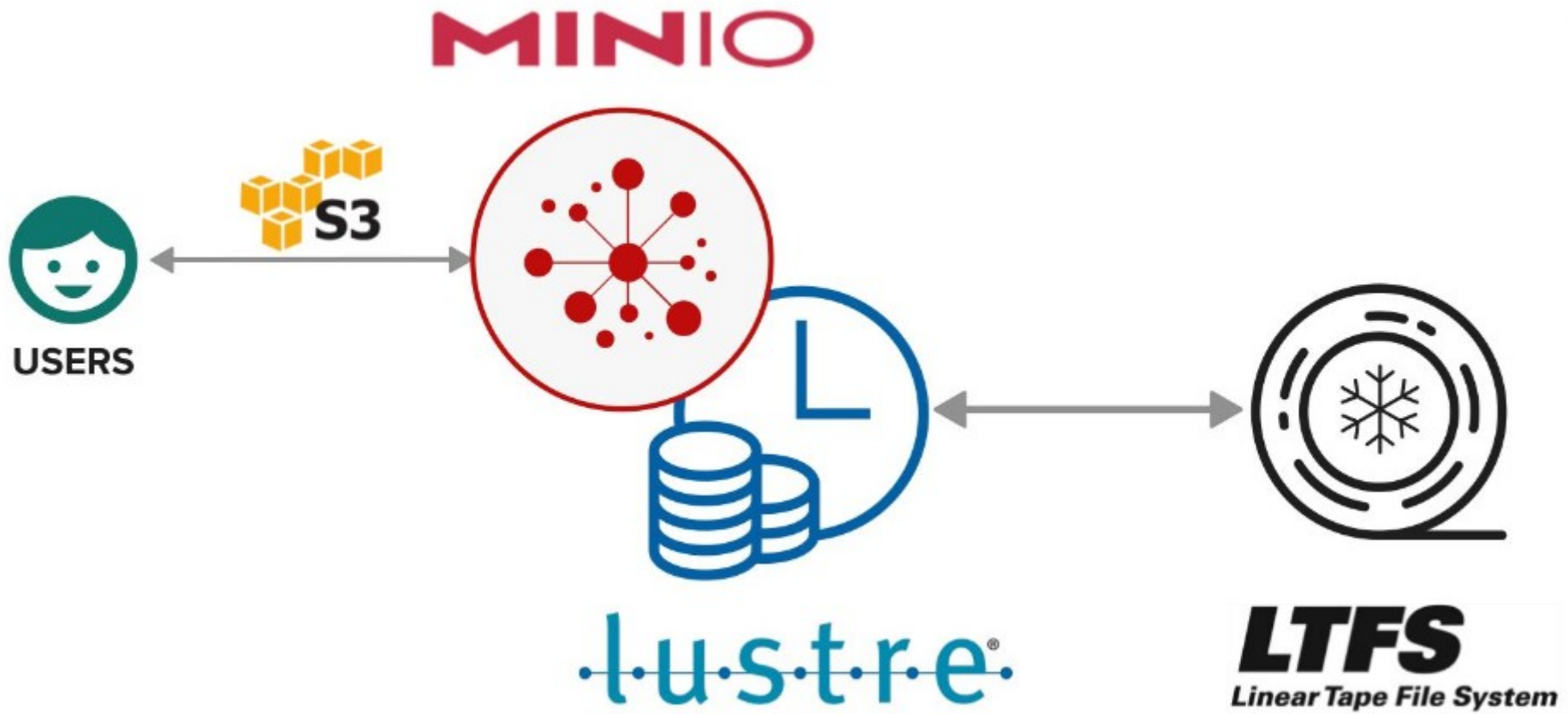


Lustre/HSM – Phobos tools chain



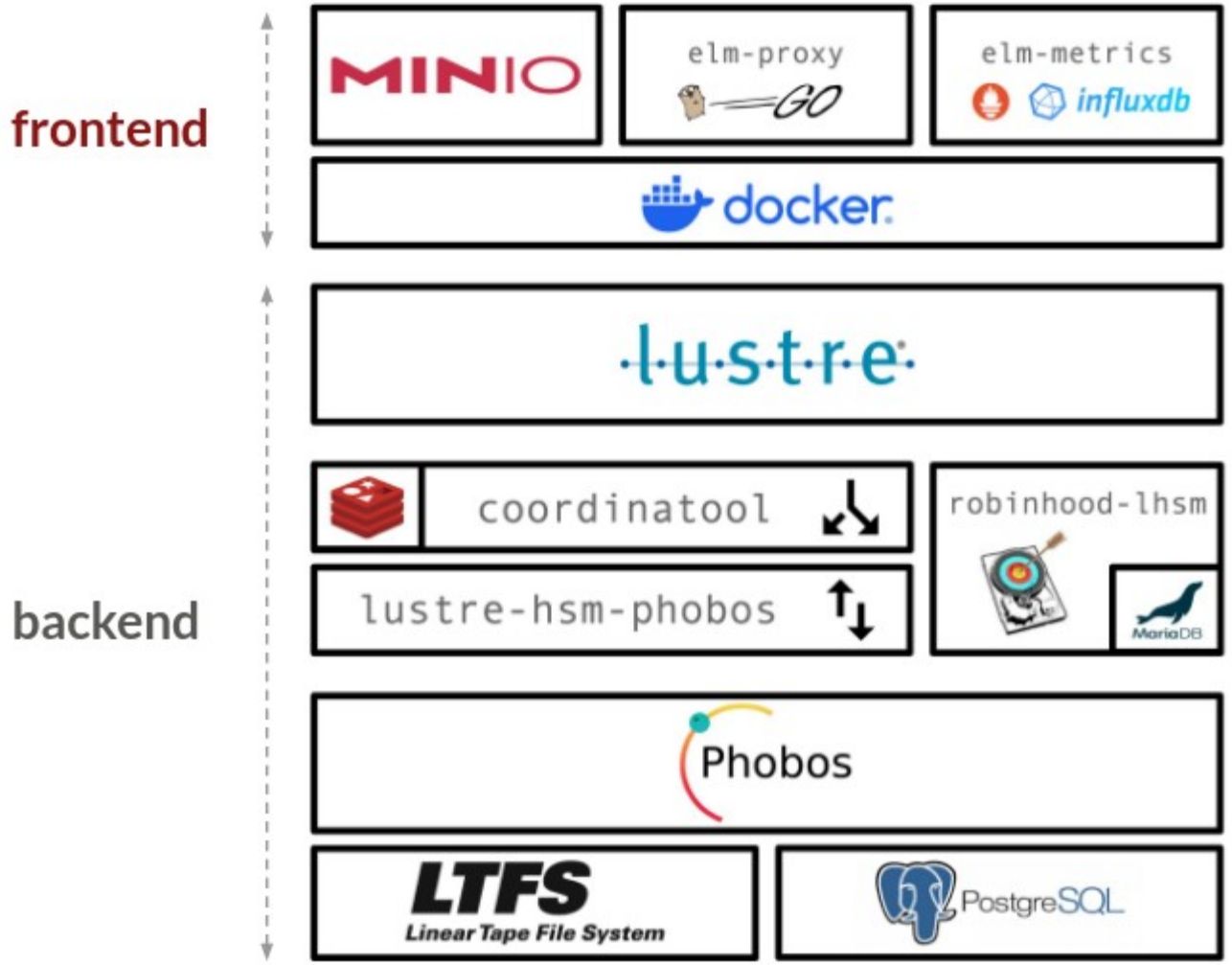
2. Production case at Stanford

Elm@Stanford : big picture



From https://www.eofs.eu/wp-content/uploads/2024/09/S_Thiell_Elm_LAD_2024.pdf

Elm@Stanford : IO stack



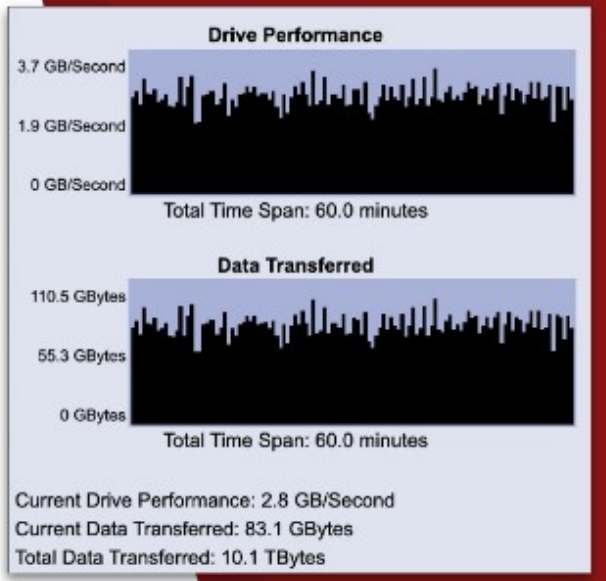
From https://www.eofs.eu/wp-content/uploads/2024/09/S_Thiell_Elm_LAD_2024.pdf

Lustre/Phobos LAD'24 "hero" run

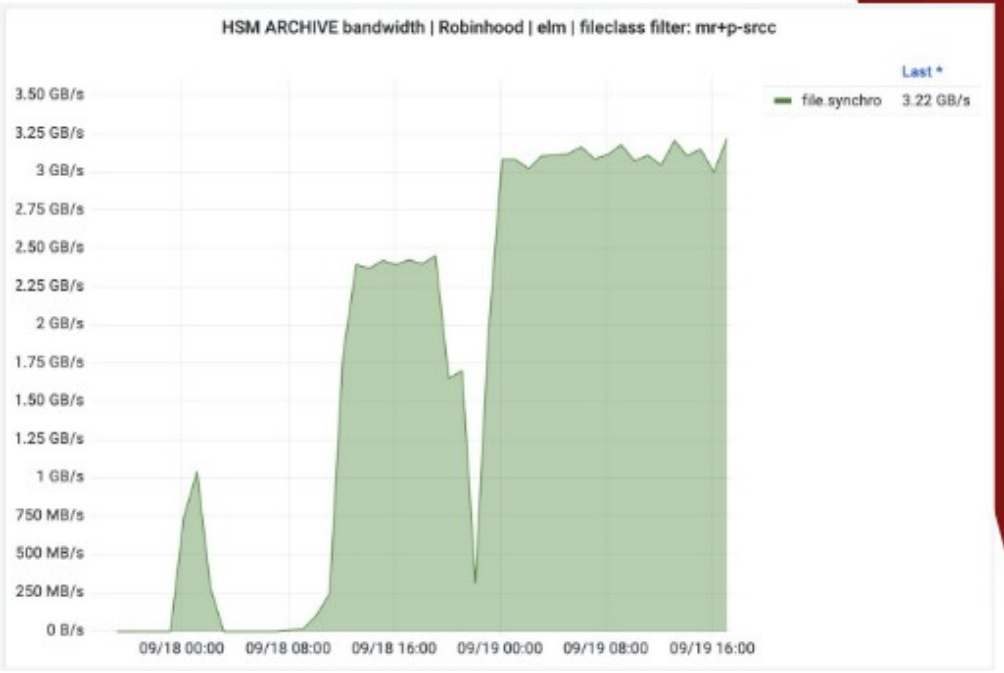
- Lustre/HSM with DNE archive run using large files, 4 data movers (one per MinIO shard), 4 LTO-9 SAS drives per mover
- Allow all drives to write:

```
# clush -w @dm phobos sched fair_share --type LTO9 --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**



not LumOS!



From https://www.eofs.eu/wp-content/uploads/2024/09/S_Thiell_Elm_LAD_2024.pdf

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 |
```

From https://www.eofs.eu/wp-content/uploads/2024/09/S_Thiell_Elm_LAD_2024.pdf

3. 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
```

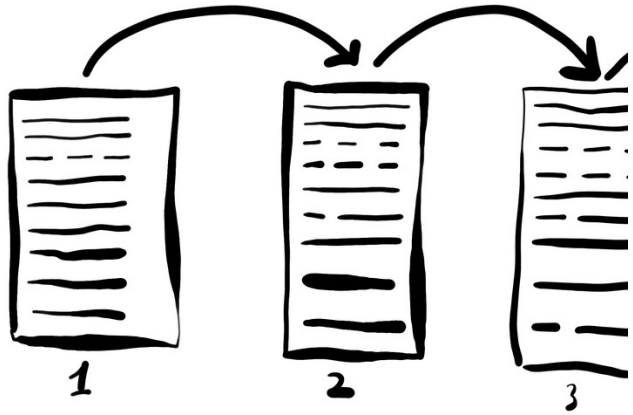
→ creates a new version of objid1

- Listing object versions

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

- Retrieving an old version

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



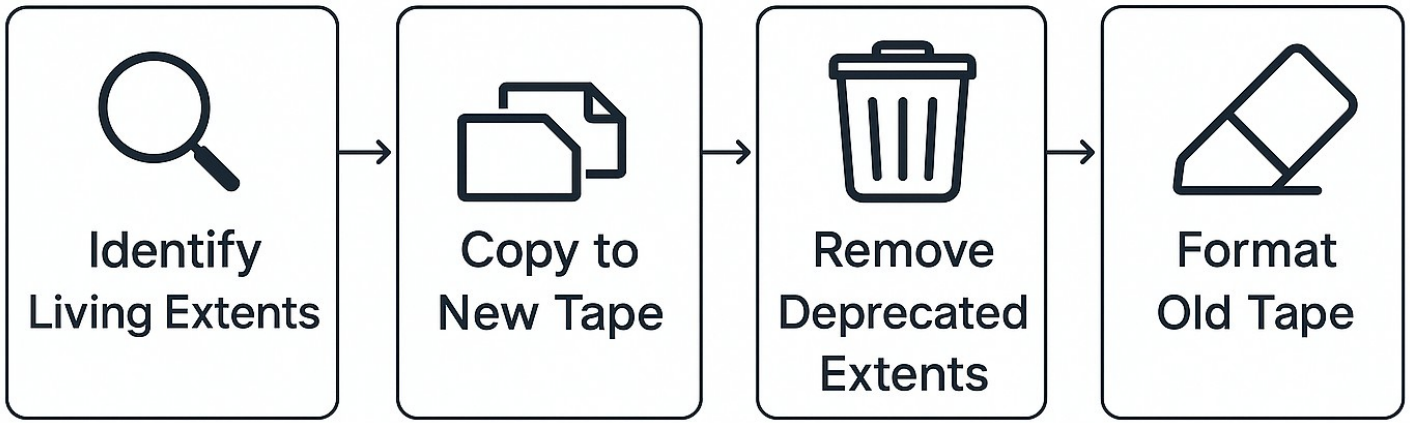
Phobos features illustrated

Repack: tape life cycle

- Optimize tape utilization
- Migrate data between tapes

```
phobos tape repack 073222L9
```

- [New in phobos 3.2] Repack ordered by offset on tape



Phobos features illustrated

Resource partitioning with media tags:

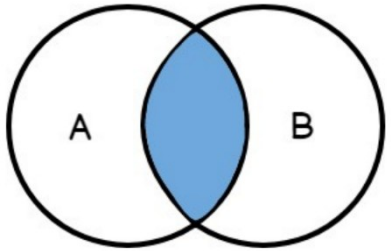
- Tagging resources

```
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

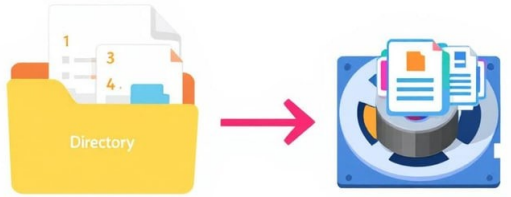
- **Grouped put:** data batches based on elapsed time, object count, or written volume, to optimize tape throughput.
- **Grouped read:** grouping read by tape to minimize tape mounts
  [New in phobos 3.2] Read ordered by offset on tape (actually: extent's creation time).

- **Dynamic and tunable fair-share** to prioritize the different type of operations (read/write/format).

- Tracking SCSI error history to identify faulty components
  (is the drive or tape faulty?)

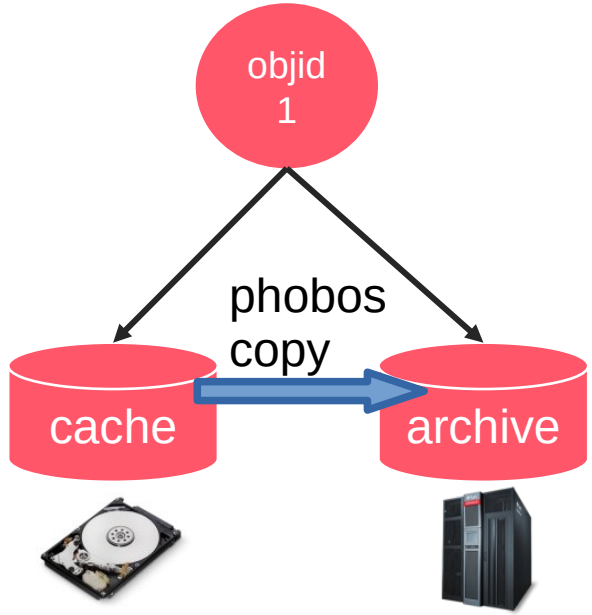
Groupings

- **Goal:** group a set of objects on the same medium, to minimize tape mounts (e.g. group object based on their source directory)
- “Groupings” are specified by an admin-defined identifier attached to objects (e.g. parent directory id).
- Media containing such objects are associated to this identifier
- For any new object with the same identifier, Phobos will try to use media associated with this identifier
- If no such media is available, a new media is associated to this identifier
- Trade-off between grouping and parallelism: configuration-driven (parameter: `fifo_max_write_per_grouping`)



Phobos internal tiering

- Phobos can manage multiple storage tiers using the “copy” feature
  - For hierarchical storage (e.g. to have a disk cache in front of tapes, or to store small files only to disks).
  - For mirroring (e.g. replication between multiple buildings)
  - For technology migration



Phobos objects copies

- **Create a copy:**

```
phobos copy create objid1 cache
```

- **Delete a copy:**

```
phobos copy delete objid1 cache
```

- **Listing copies:**

```
phobos copy list objid1
```

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

Phobos objects copies

Option --copy-name 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 associated with put parameters (layouts, family, …)

Phobos objects copies

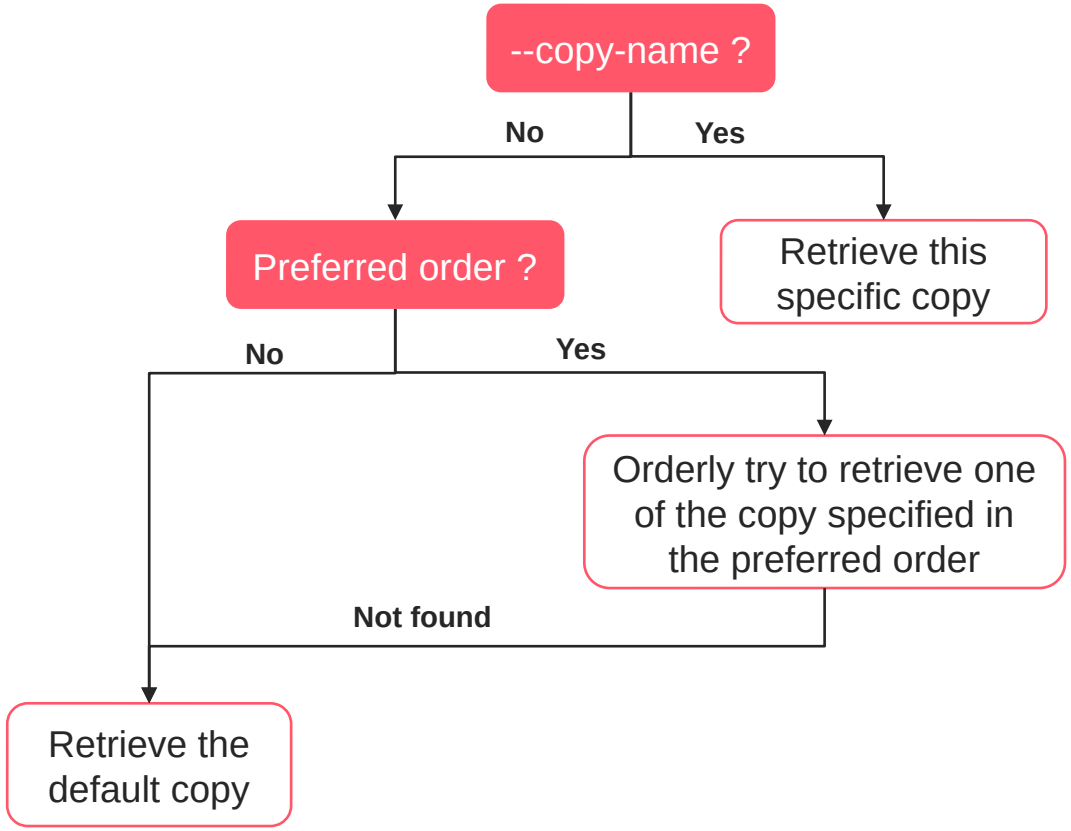
Get preferred order

- 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

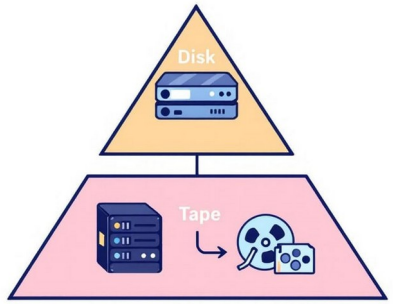
- Can be set in the configuration



4. New features update

Internal HSM (Phobos 3.1)

- Enable automated data movement between phobos copies
- Available commands :
  - phobos_hsm_sync_dir
  - phobos_hsm_release_dir
- Sync:
  - Inputs : source copy name, target copy name, time frame
  - Synchronise all eligible copies from the local host
- Release:
  - Same parameters + based on high/low thresholds
  - Purge all eligible entries from the local host



Metrology (Phobos 3.3)

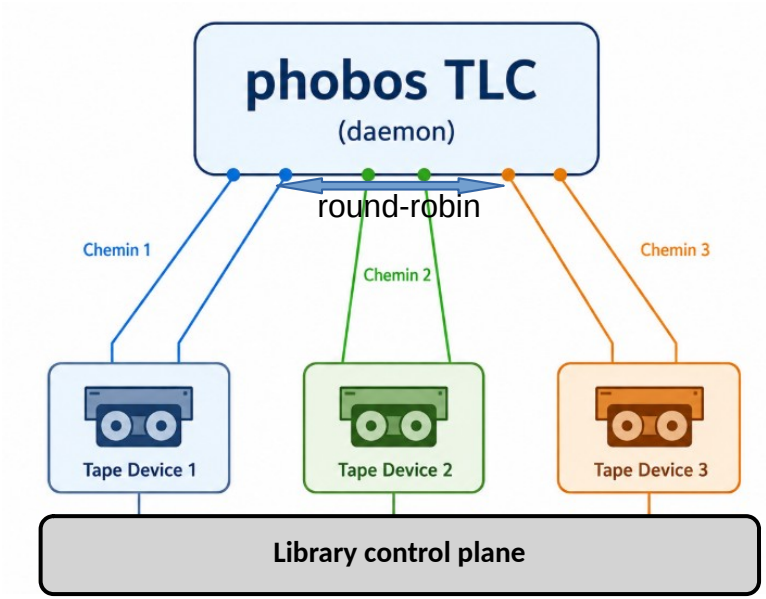
- Provides statistics about Phobos functioning:
  - Stats per daemon (phobosd, TLC – Tape Library Controller)
  - Requests stats per media family, per device, wait queues…
    - See: https://github.com/cea-hpc/phobos/blob/master/doc/stats.md
- Monitoring tool friendly format

```
$ phobos stats

action=read                              req.count=212
action=write                             req.count=1537
family=tape                              sched.incoming_qsize=7
family=tape,device=03FA5328              dev.nb_mount=35
family=tape,device=03FA5328              dev.nb_umount=34
family=tape,device=03FA5328              dev.requested_sync=546
family=tape,device=03FA5328              dev.effective_sync=52
family=tape,device=03FA5328              dev.tosync_size=1546987640
...
```

Multi-path TLC (Phobos 3.3)

- **Goal:** use multiple paths to send library control commands
- List of usable devices specified in the configuration
- SCSI Requests are sent to the devices with a Round-Robin strategy
- If SCSI ping fails, the device is blacklisted
  - Possible reset using « phobos tlc --refresh »



Contribution to MHVTL

- A VTL is very convenient for testing Phobos intensively
- We previously used Quadstor VTL, but it’s no longer freely available
- MHVTL: open-source Virtual Tape Library
- **CEA contribution**:
  - Enabling the support of tape partitioning in MHVTL

    → Enables the use of LTFS with MHVTL
- Contribution integrated as major release 1.8.0

5. Roadmap

Upcoming features (2026-2027)

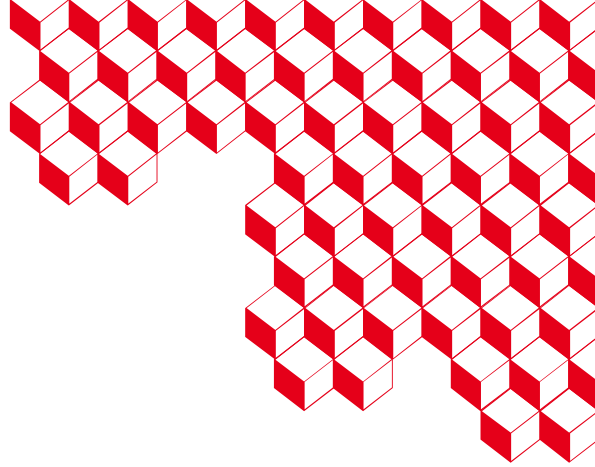
- Automated rebuild of damaged objects
- Full validation of Phobos Disaster Recovery Plan
- Query objects by regexp on metadata values
- Concurrent I/Os on disk devices
- Repack to non-empty tape
- Advanced media stats to drive repack decisions
- Advanced garbage collection of object versions (smart repack)
- Media auto-tag
- Internal HSM optimisation

Test, Use & Contribute

Interested?

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

- Commercial support available at: https://ayudata.fr



Thank you !

To contact the team: st-hpc@cea.fr