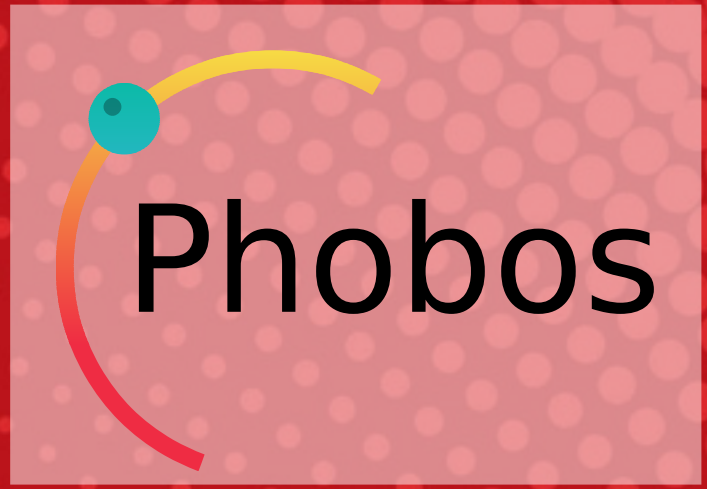
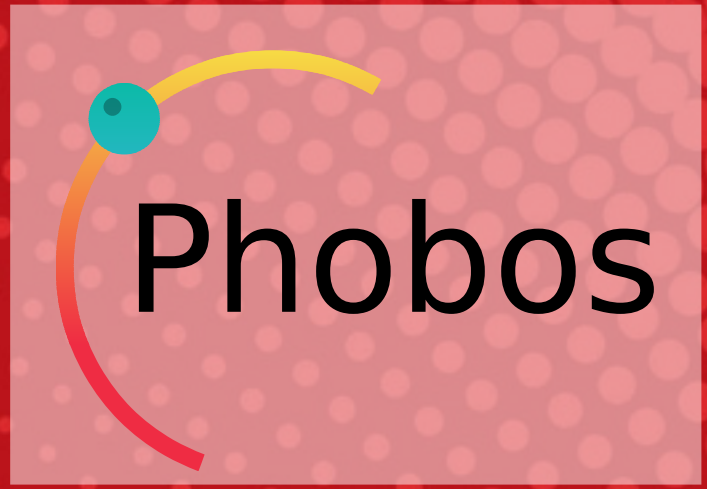
HPC-IODC 2020, HPC I/O in the Data Center Workshop, 06/25/2020



Phobos: an open-source object store implementing tape library support

Patrice LUCAS, patrice.lucas@cea.fr

HPC-IODC 2020, HPC I/O in the Data Center Workshop, 06/25/2020



Phobos: an open-source object store implementing tape library support

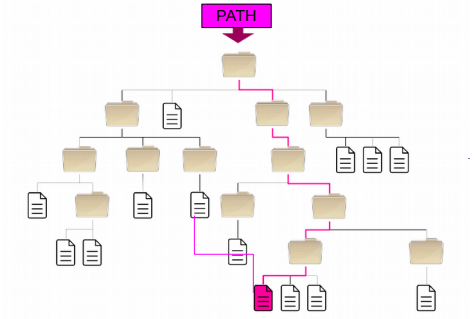
Patrice LUCAS, patrice.lucas@cea.fr

Next scale of mass storage

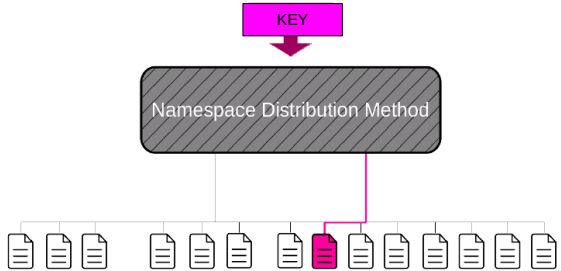
- Exaflopic supercomputers in the 2020's
- Huge amounts of data to ingest: petabytes per day
- Huge amounts of data to store: exabytes

	Today	Tomorrow
Daily production	Hundreds of TB	Petabytes
Storage system capacity	Hundreds of PB	Exabytes

Suppressing POSIX filesystem's bottlenecks



Addressing entries in a POSIX file system

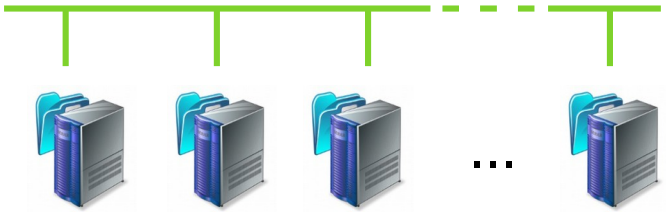


Addressing entries in an object-store

- Object stores have proved their scalability
- Widely adopted for Internet services, Cloud computing, social networks...

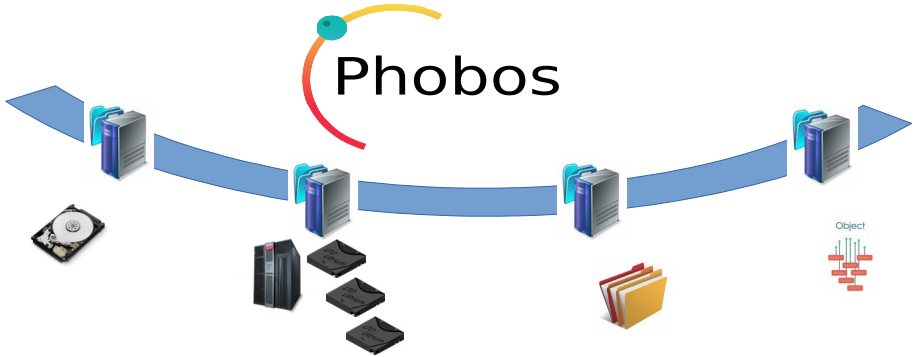
Needs for extremely scalable storage systems at a reasonable price

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



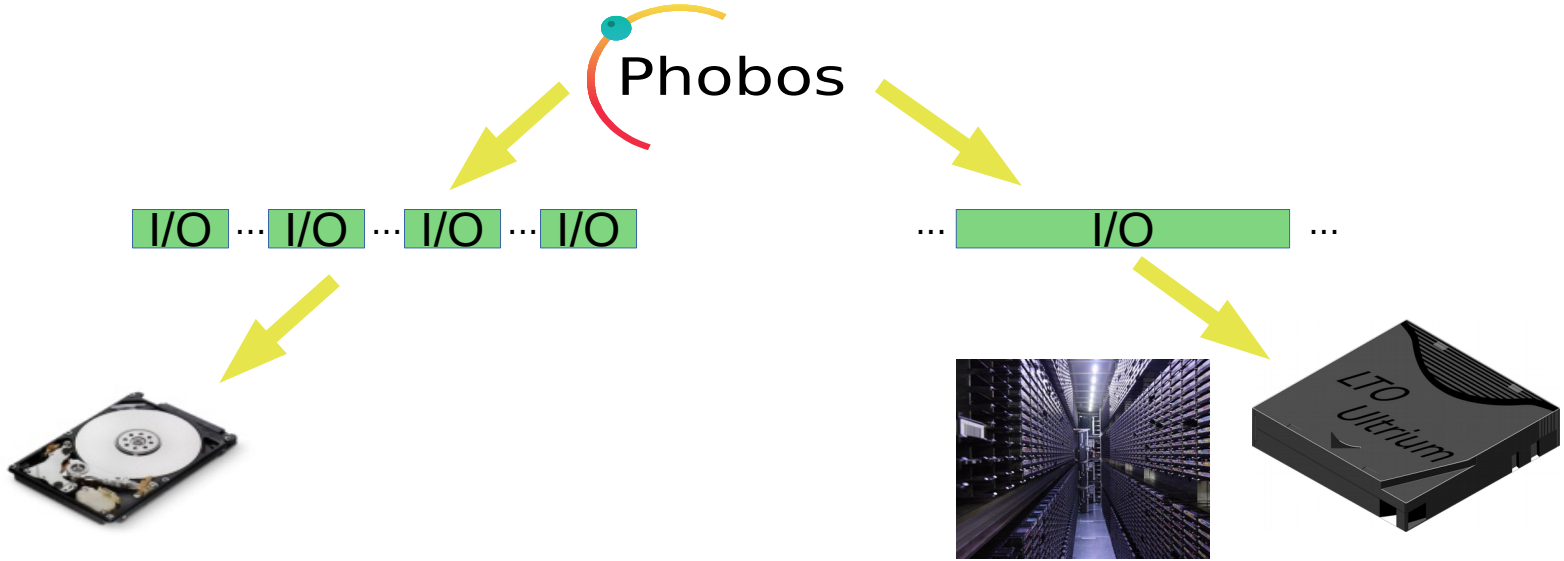
Phobos: Parallel Heterogeneous Object Store

- Manages a distributed set of file systems on different storage media technologies
  - hard drives or magnetic tapes



Phobos: optimizes I/O access depending on the storage technology

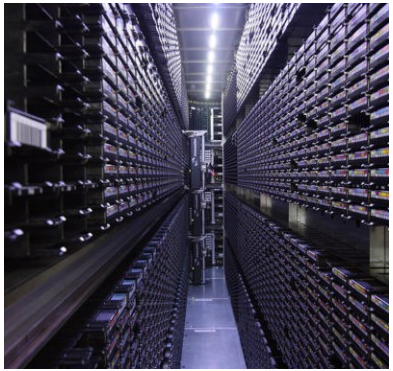
- minimizing data syncs for magnetic tapes



Phobos: directly manages a tape library (SCSI)

- Supports all common tape drives model

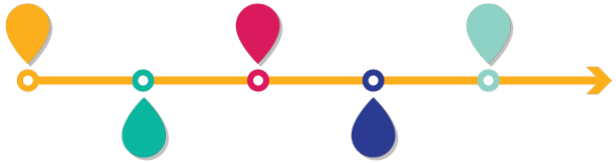
Phobos



Design guidelines

- Scalability and fault-tolerance
- Based on open formats, open protocols, interoperable
  - E.g. LTFS as tape filesystem (ISO/IEC 20919:2016)
- Simple and common interfaces (REST, object stores API)
- Simple administration (intuitive, admin-friendly CLI)
- Light, easy to deploy, easy to maintain
  - As of today: 14k lines of C, 2.5k lines of Python

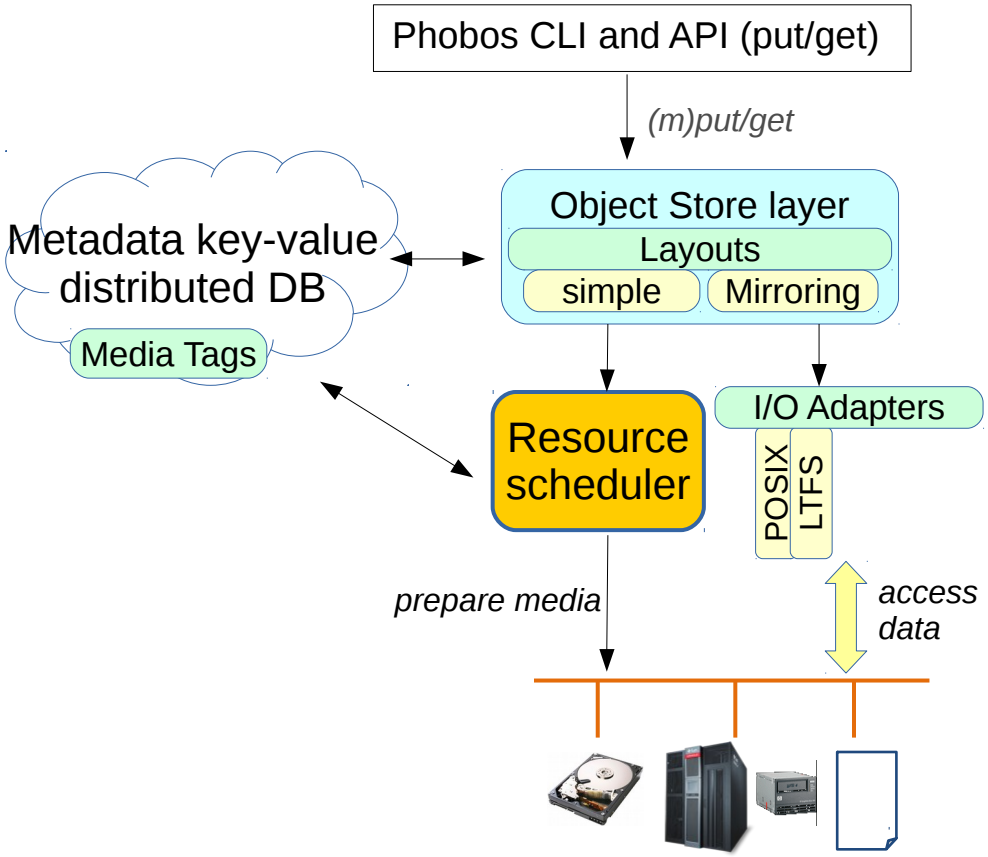
History of the project



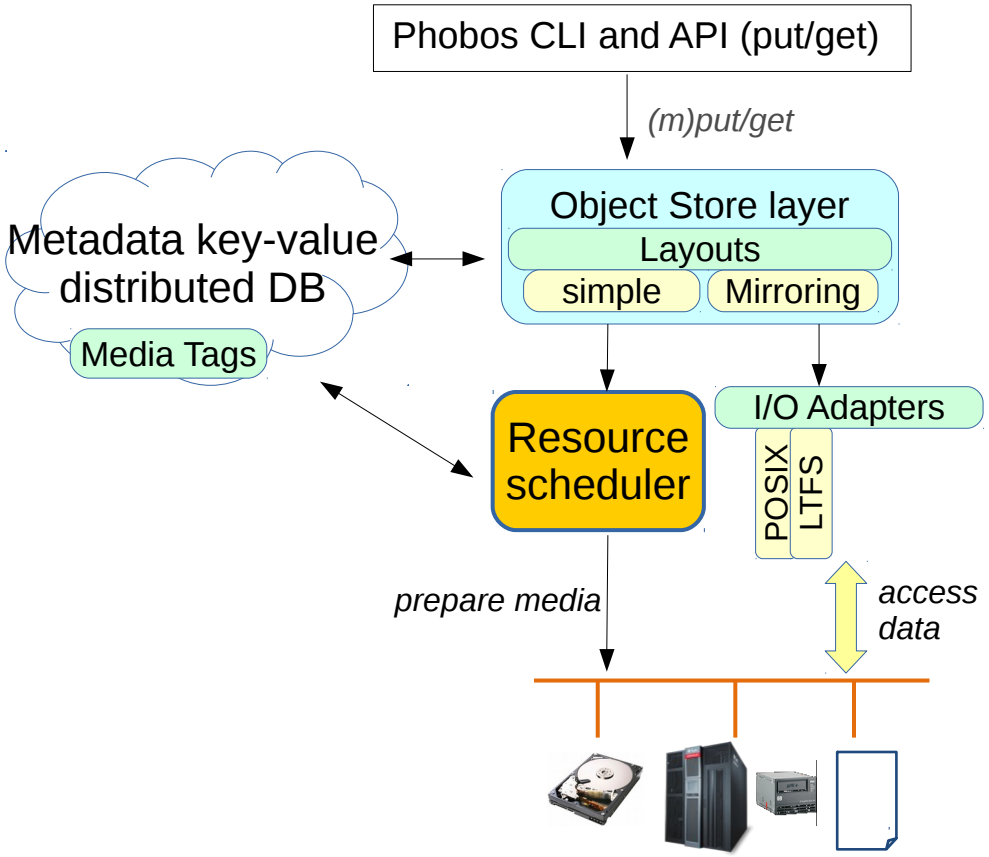
- 2013: high level design
- 2014-2015: development of the initial version
  Scope:
  - Storage on tape, or in a filesystem
  - SCSI-controlled tape library and LTO drives
  - Single server
- 2016: Phobos v1.0 in production
  - Multi-Petabyte storage of genomics data
  - IBM TS3500 library, LTO5/6 drives

FRANCE GÉNOMIQUE

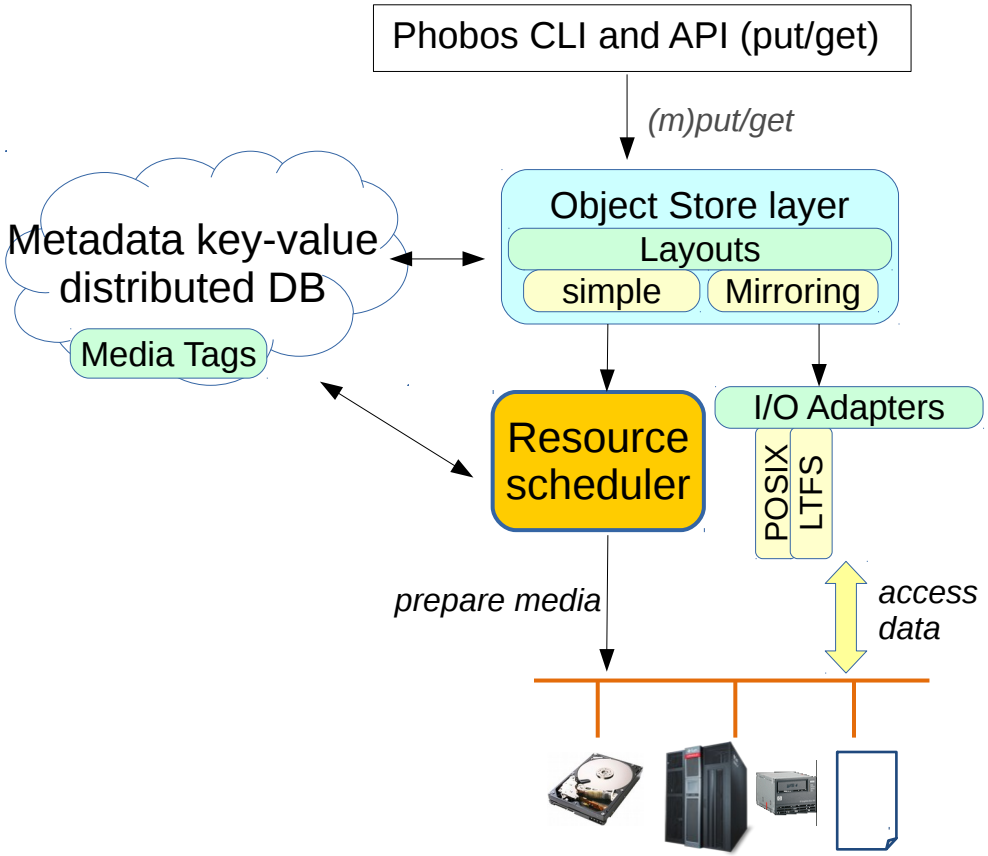
- 2019: Phobos made available on github as open-source (LGPL v2.1)
- 2020: Implementation of an S3 front-end (collab. with ICHEC and DDN) + required features in the Phobos core
- Next steps:
  - Advanced IO scheduling
  - Parallelization across multiple servers


Phobos CLI and API (put/get)
(m)put/get
Metadata key-value distributed DB
Media Tags
Object Store layer
Layouts
simple
Mirroring
Resource scheduler
I/O Adapters
POSIX
LTFS
prepare media
access data

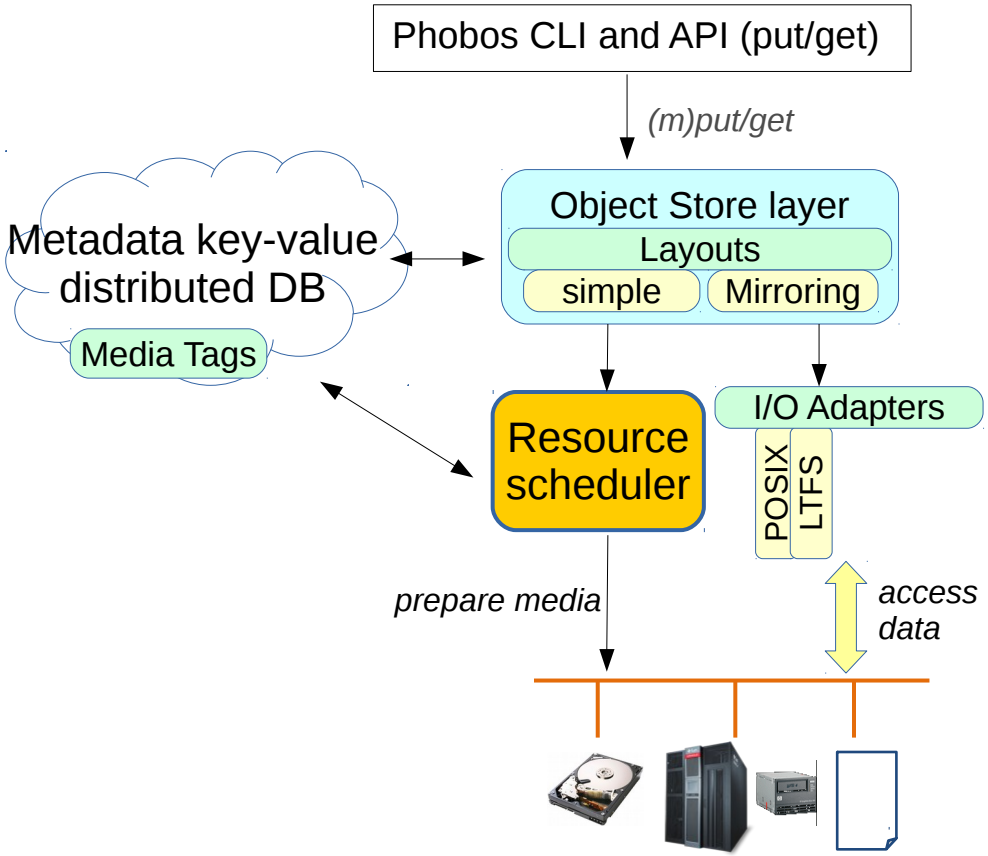
- IO adapters: multiple storage technologies (Posix, LTFS)



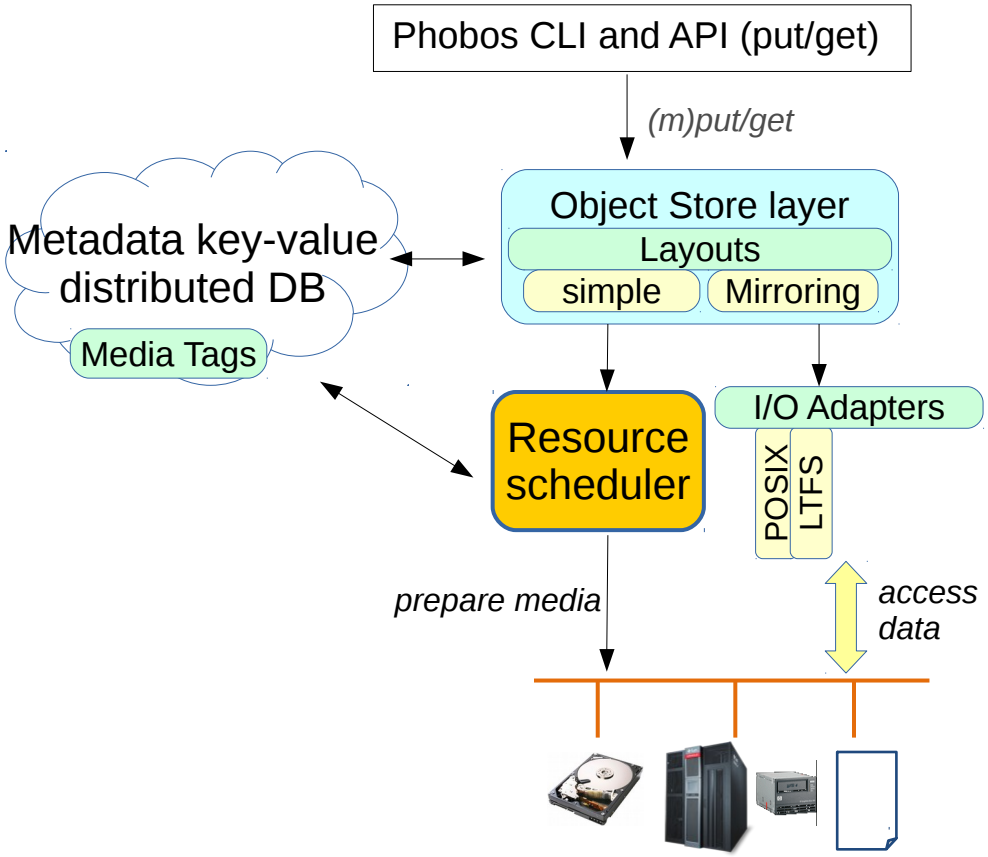
- IO adapters: multiple storage technologies (Posix, LTFS)
- Layout plugins: performance and fault-tolerance



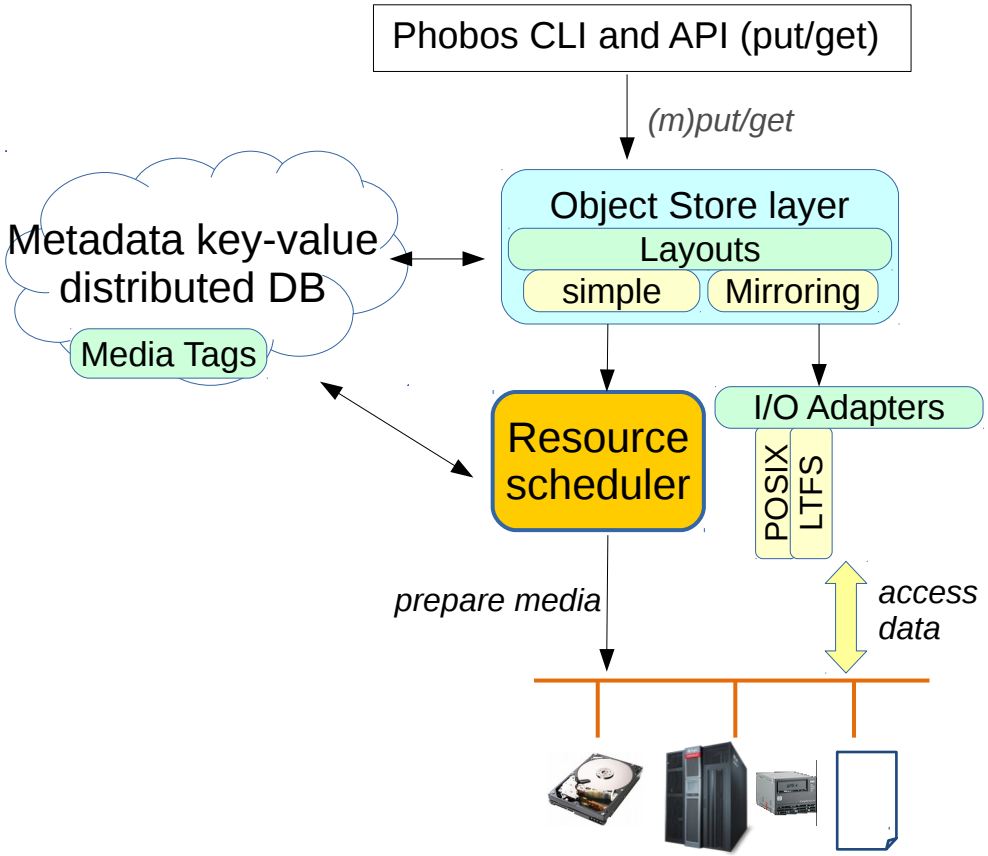
- IO adapters: multiple storage technologies (Posix, LTFS)
- Layout plugins: performance and fault-tolerance
- Tags: storage partitioning

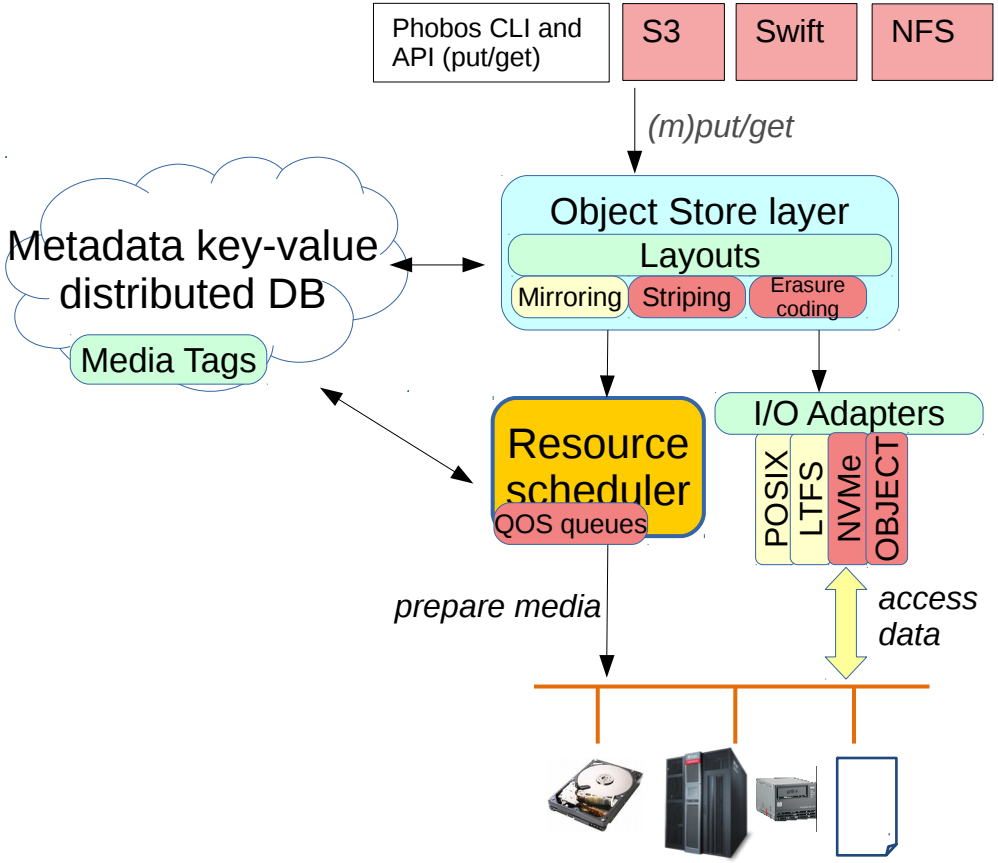


- IO adapters: multiple storage technologies (Posix, LTFS)
- Layout plugins: performance and fault-tolerance
- Tags: storage partitioning
- Resource scheduling: optimizes tape fill rate, minimizes tapes mounts

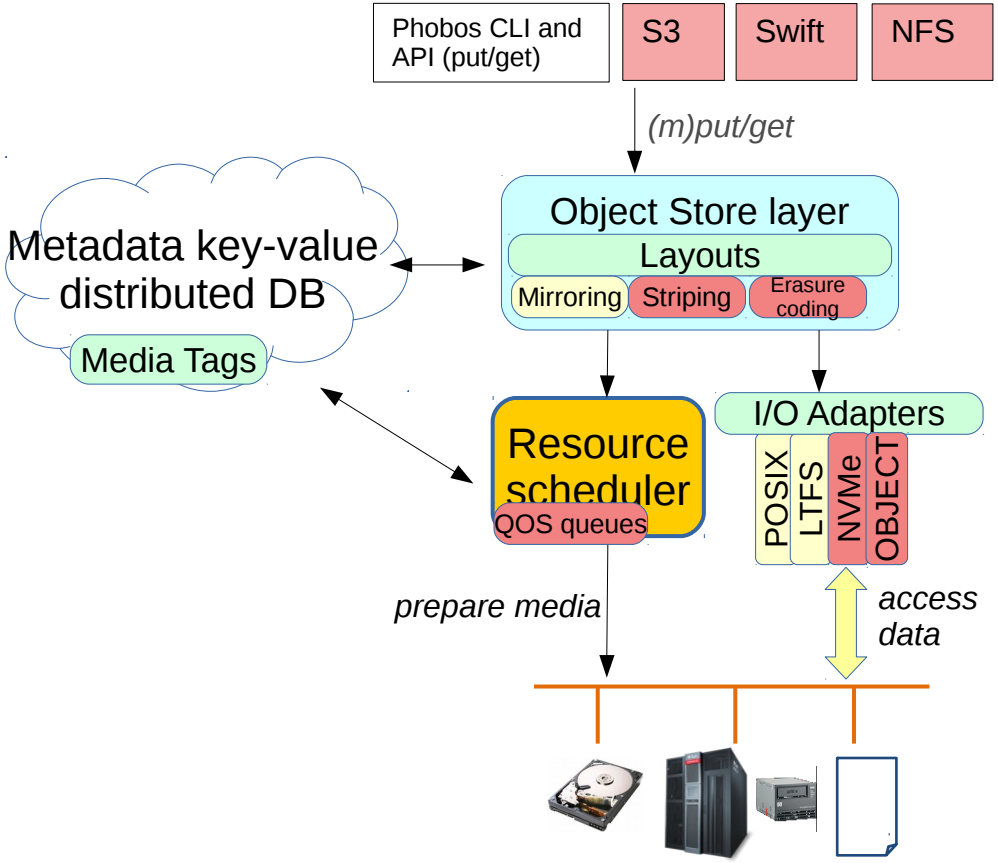


- IO adapters: multiple storage technologies (Posix, LTFS)
- Layout plugins: performance and fault-tolerance
- Tags: storage partitioning
- Resource scheduling: optimizes tape fill rate, minimizes tapes mounts
- Key-value metadata schema:
  - Distributed NoSQL Database
  - Saved within objects on media (recovery, tape import)

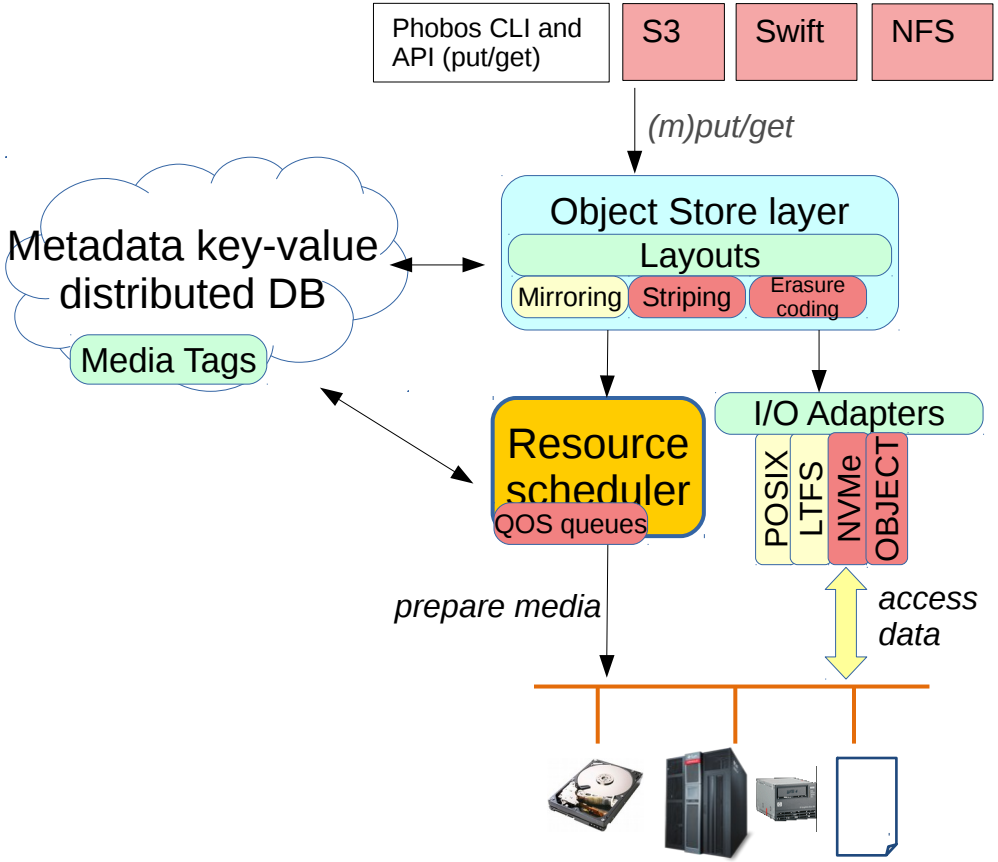



Phobos CLI and API (put/get)
S3
Swift
NFS
(m)put/get
Metadata key-value distributed DB
Media Tags
Object Store layer
Layouts
Mirroring
Striping
Erasure coding
Resource scheduler
QOS queues
I/O Adapters
POSIX
LTFS
NVMe
OBJECT
prepare media
access data

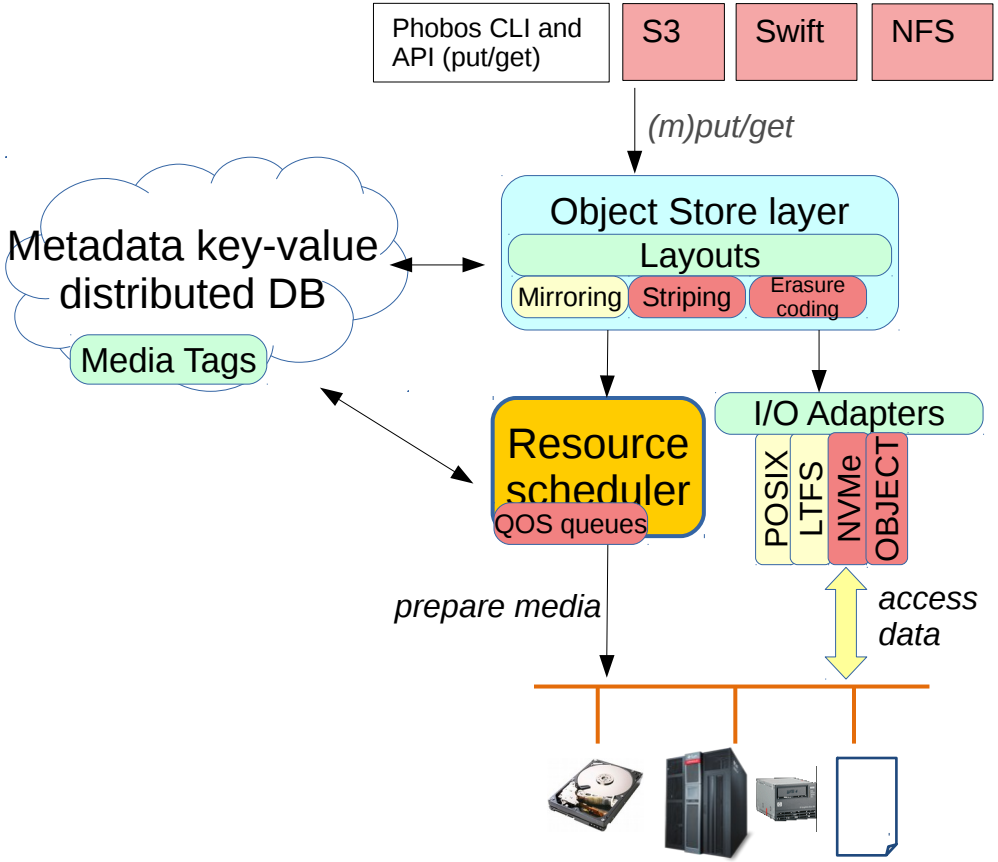
- Providing S3, Swift and NFS connectors



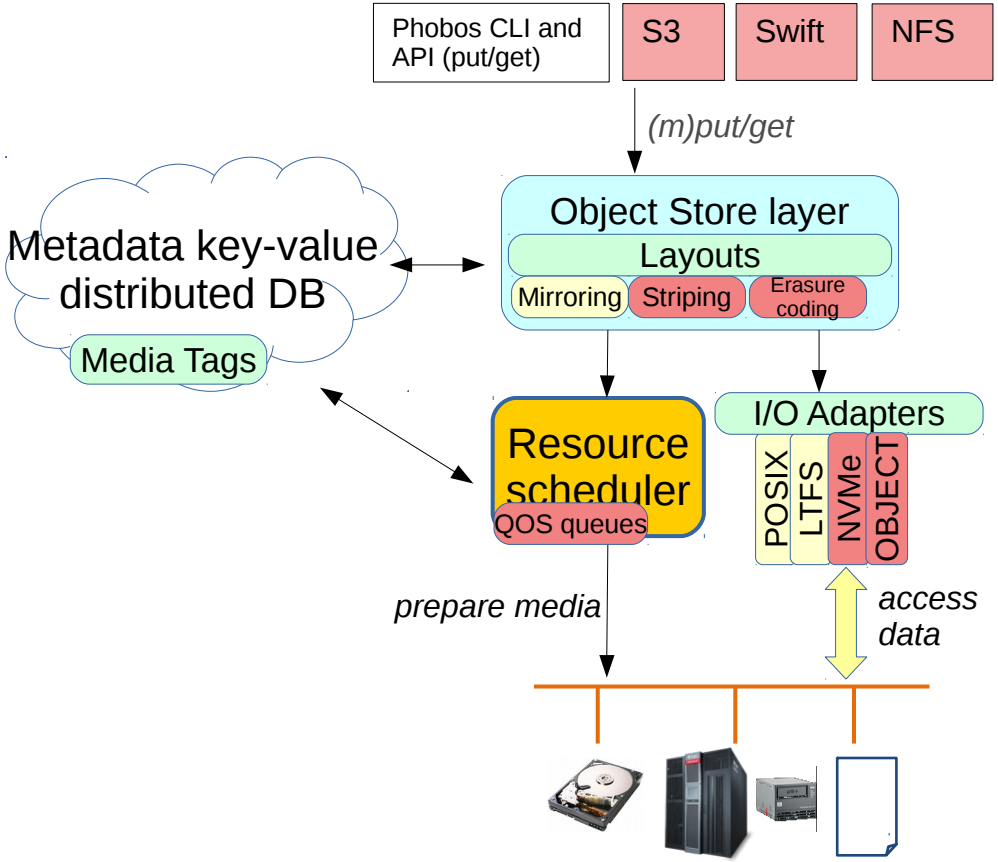
- Providing S3, Swift and NFS connectors
- Adding new layouts: striping, erasure-coding



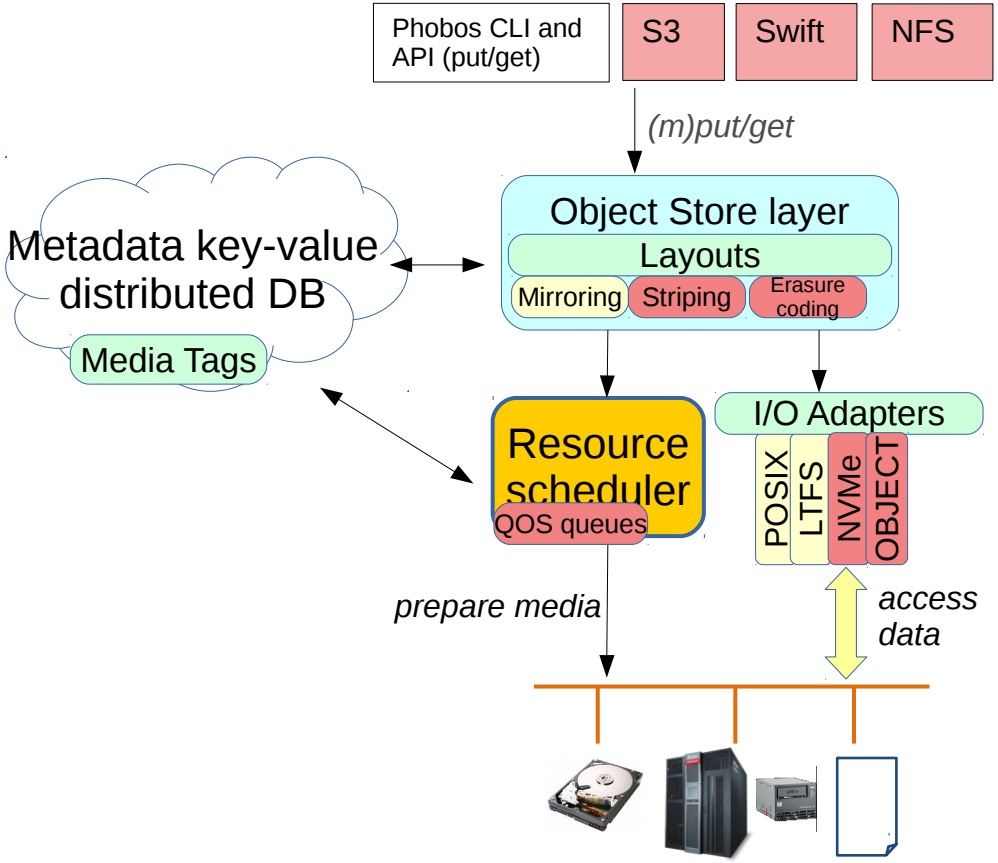
- Providing S3, Swift and NFS connectors
- Adding new layouts: striping, erasure-coding
- New IO adapters: NVMe, Object

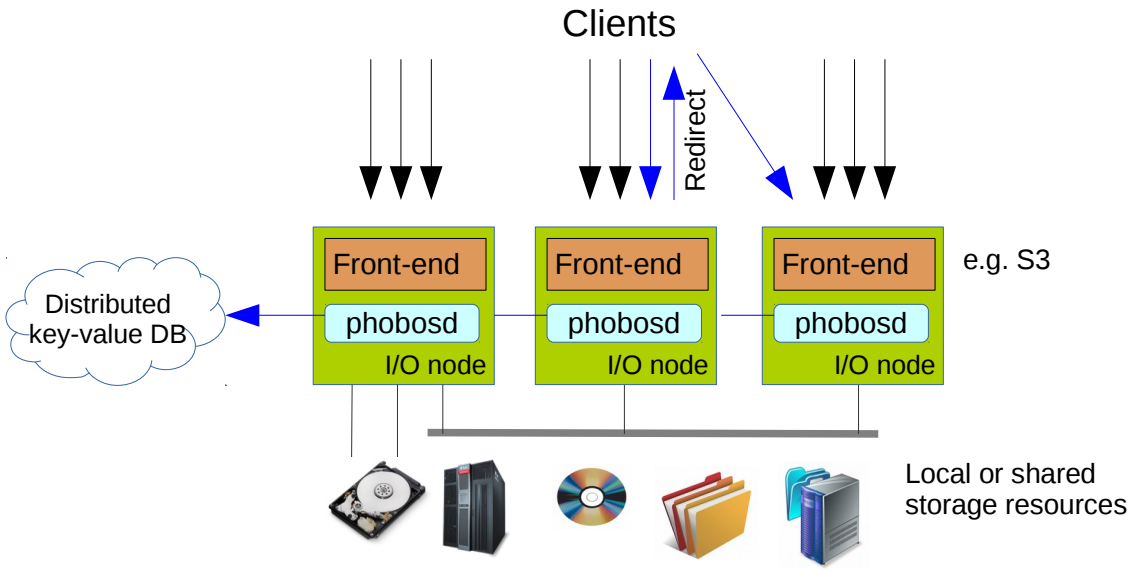


- Providing S3, Swift and NFS connectors
- Adding new layouts: striping, erasure-coding
- New IO adapters: NVMe, Object
- Media life cycle management: automatic migrations between storage technologies



- Providing S3, Swift and NFS connectors
- Adding new layouts: striping, erasure-coding
- New IO adapters: NVMe, Object
- Media life cycle management: automatic migrations between storage technologies
- Optimizing resource scheduler policies: prioritizing and grouping I/O





- Synchronization in the distributed mode:
  - Through the distributed key-value DB (object location, resource reservation…)
  - Redirection of client requests to the preferred I/O node (max 1 hop)

Setting up a tape storage in a couple of commands

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

phobos tape add -t lto6 [073200-073222]L6

phobos tape format --unlock [073200-073222]L6
```

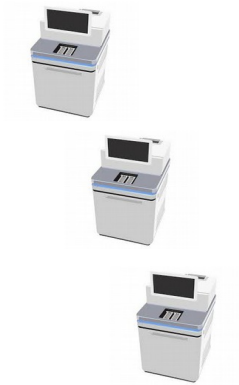
That's done! Your system is ready for I/Os.

Example of use-cases

- Multi-petabyte genomics datasets
- In production since 2016

FRANCE GÉNOMIQUE

DNA sequencers



Phobos

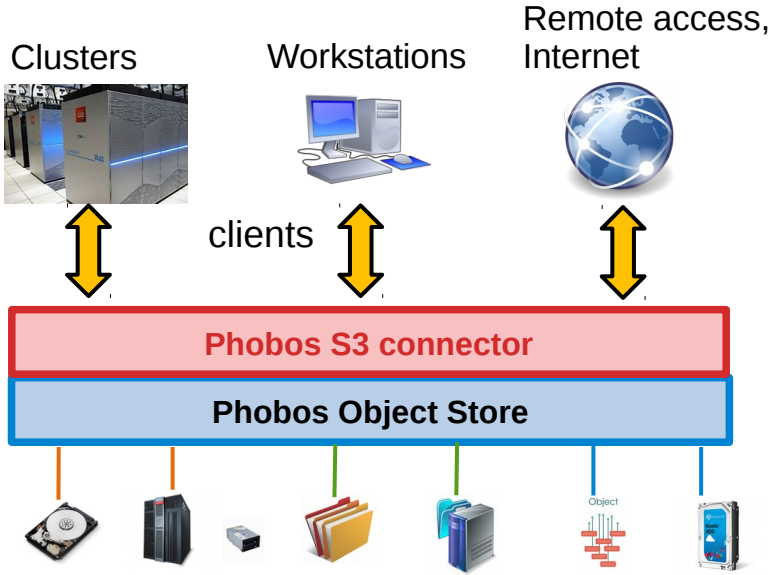
- IBM TS3500 tape library (SCSI)
- LTO6 and LTO8 drives

HPC data clusters

TGCC
Très Grand Centre de calcul du CEA

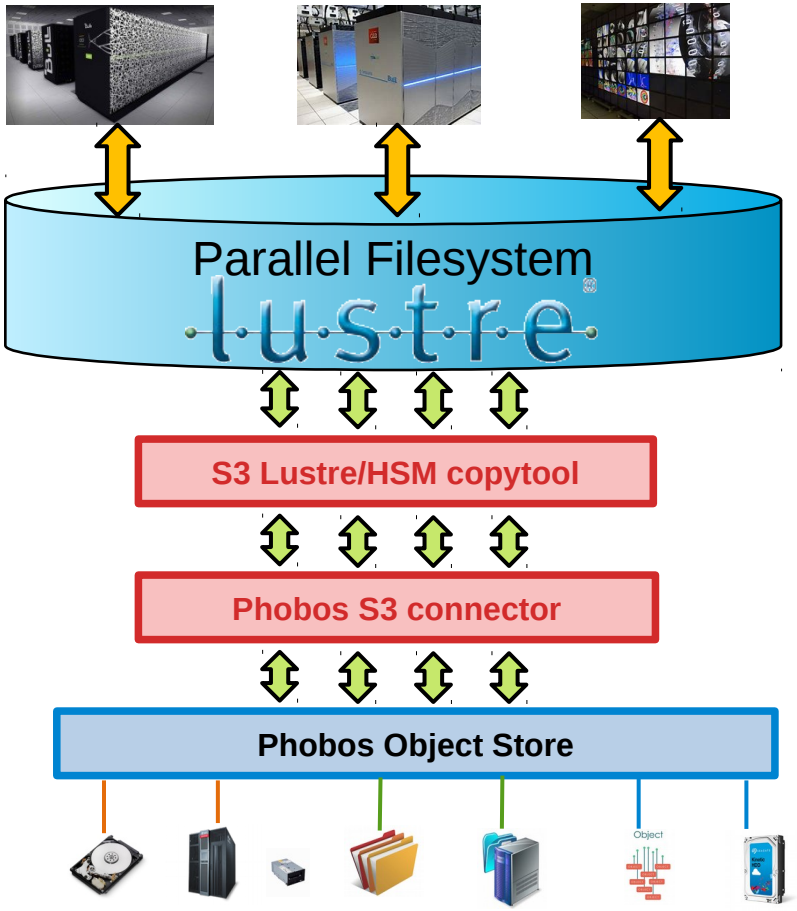
Object store with an S3 interface

- S3 interface exposed to end-users
- Phobos: high-performance, scalable storage
  - Can manage a wide variety of capacitive storage, including tape libraries
  - Provides an easy/uniform management of these storages



Phobos as a Lustre/HSM backend

- Lustre: filesystem user front-end
- Phobos as capacitive and scalable backend (hierarchical storage)



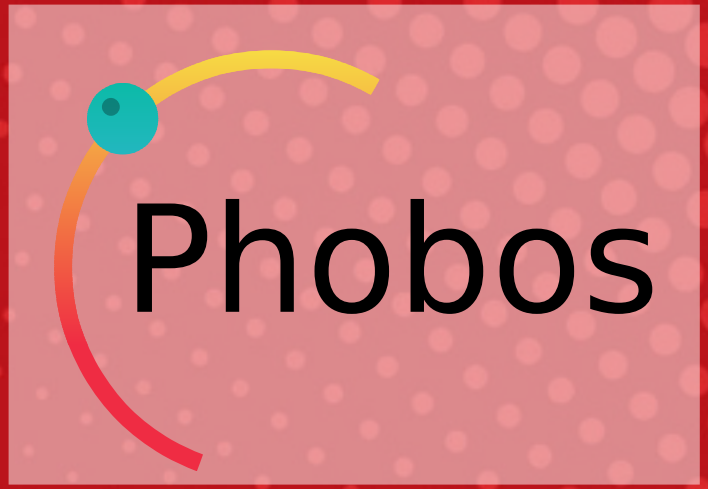
Summary

- Tape object storage at scale (and more)
- Phobos is open-source, available on github:

  https://github.com/cea-hpc/phobos

- Contributions are welcome, as well as testers!

DE LA RECHERCHE À L'INDUSTRIE



https://github.com/cea-hpc/phobos

Thank you for your attention!

Patrice LUCAS, patrice.lucas@cea.fr