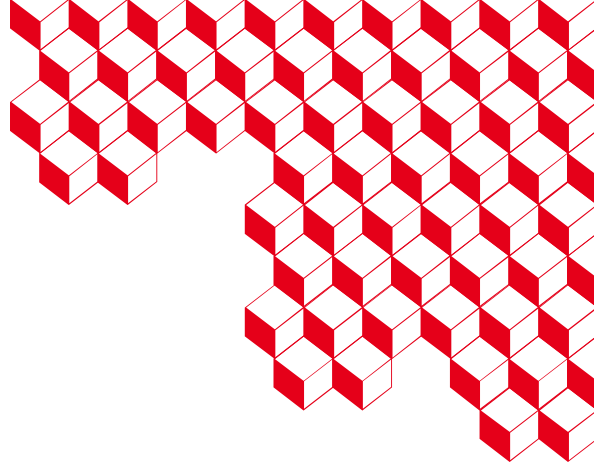


RobinHood v4 Progress Report

LAD'24 – 23rd of September, 2024

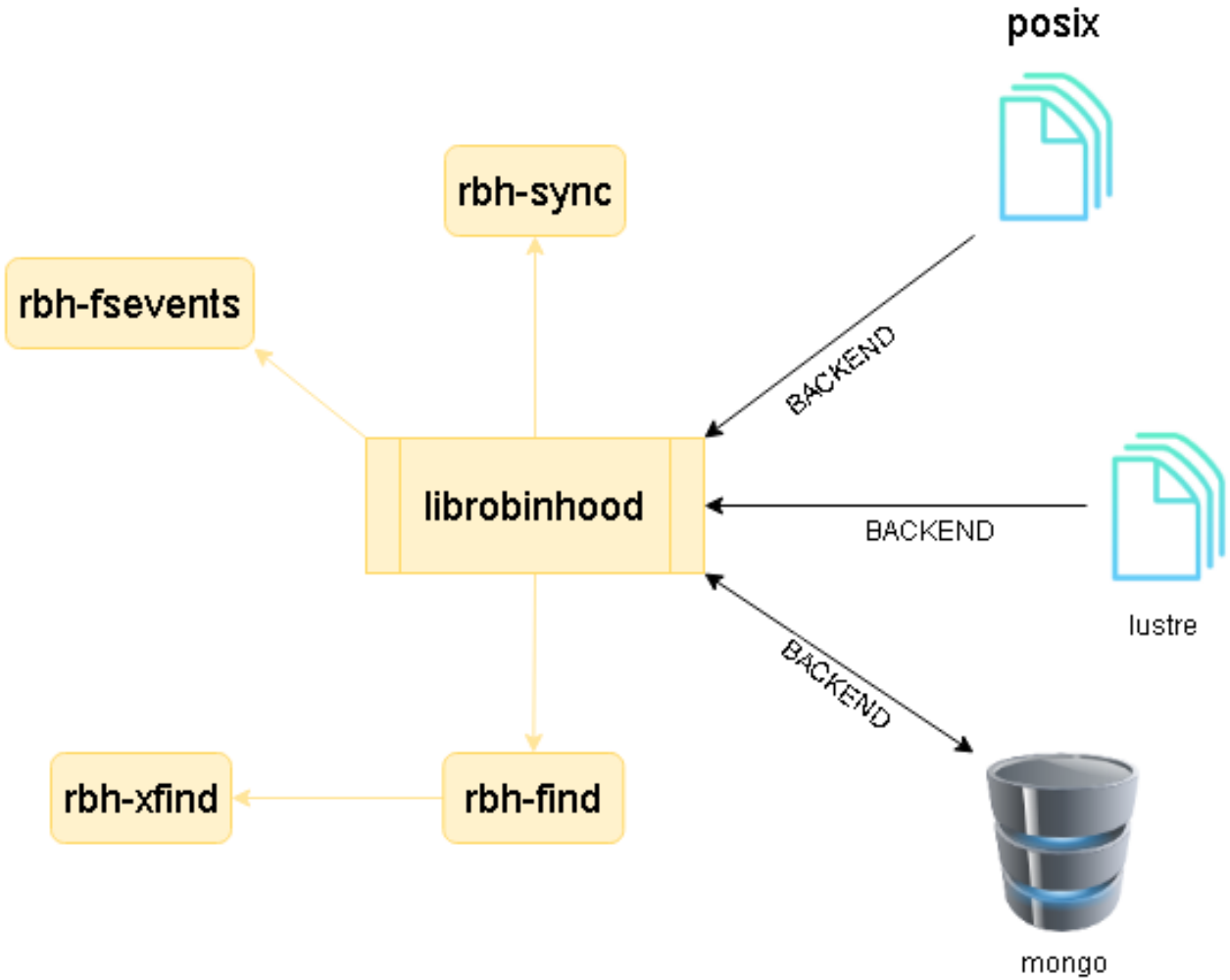
Yoann VALERI, yoann.valeri@cea.fr

Focus on RobinHood

Providing efficient and easy to use means to replicate and query any storage system's metadata

Focus on RobinHood (as of last year)

Tool suite centered around a *core* library using a NoSQL database



What's new since last year? (1)

- **librobinhood**:
  - Error skiping
  - Configuration
    - By a configuration file or environment variables
    - Management of extended attributes types, database URL, …

```
rbh-sync --no-skip rbh:posix:test_directory rbh:mongo:test_db
```

```
---
retention_xattr: "user.expires"
mongodb_address: "mongodb://localhost:27017"
xattrs_map:
    user.blob_int32: int32
    user.blob_int64: int64
    user.blob_uint32: unsigned int32
    user.blob_uint64: unsigned int64
    user.blob_string: string
    user.blob_boolean: boolean
---
```

What's new since last year? (2)

- **librobinhood**:
  - New backends:
    - Hestia backend for the IO-SEA European project
    - MPIFileUtils backends → Can be used for POSIX and Lustre
    - MPIFile backend
      - Use MPIFileUtils tools with RobinHood → dwalk, dfind, …

```
rbh-sync rbh:hestia: rbh:mpi-file:test_mpi_file
mpirun -np 4 rbh-sync rbh:posix-mpi:first_directory rbh:mongo:test_db
mpirun -np 4 rbh-sync rbh:lustre-mpi:second_directory rbh:mongo:test_db
rbh-sync rbh:mongo:test_db rbh:mpi-file:test_mpi_file
```

```
mpirun -np 128 dfind -i test_mpi_file -v -o output_file
```

What's new since last year? (3)

- **rbh-find**:

```
rbh-find rbh:mongo:test_db -exec grep -H "test data" {} ";"
rbh-find rbh:mongo:test_db -size +3G -delete
```

  - « -printf »
  - « -exec » and « -delete »
- **rbh-lfind**:
  - New filters: Component start, pool name, layout pattern, stripe count, stripe size
  - Better handling of the retention attributes with « -printf »
    - Printing of the attribute set by the user and the expiration date computed

```
rbh-lfind rbh:mongo:test_db -pool "to_archive" -printf "%p %e %E\n"
```

What's new since last year? (4)

- **rbh-fsevents**:
  - Deduplication:
    - parameter for the number of entries to keep in memory for deduplication
    - Prevent unnecessary interaction with the mirror system
  - Acknowledgement
  - Changelog dumping → keep a permanent record of the changelogs

```
rbh-fsevents src:lustre:lustre-MDT0000?ack-user=cl4 -
rbh-fsevents --batch-size 500 src:lustre:lustre-MDT0000 -
```

```
rbh-fsevents --dump "-" --enrich rbh:lustre:/mnt/lustre/test_dir src:lustre:lustre-MDT0000 rbh:mongo:test_db

lustre-MDT0000: 717 01CREAT 1726132808.352593033 0x0 t=[0x200000404:0x14b:0x0] p=[0x200000404:0x14a:0x0] test_file J=touch.0
lustre-MDT0000: 718 11CLOSE 1726132808.352956554 0x42 t=[0x200000404:0x14b:0x0] J=touch.0
lustre-MDT0000: 719 06UNLNK 1726132808.353946316 0x1 t=[0x200000404:0x14b:0x0] p=[0x200000404:0x14a:0x0] test_file J=rm.0
```

What's new since last year? (5)

- **rbh-capabilities**:
  - Get the list of installed backends
  - Get the capabilities of a backend
  - Ease the use of RobinHood v4

```
rbh-capabilities --list
List of installed backends:
- mongo
- posix
- lustre

rbh-capabilities mongo
Capabilities of mongo:
- filter: rbh-find [source]
- synchronisation: rbh-sync [source]
- update: rbh-sync [source]
- branch: rbh-sync [source for partial processing]
```

What's new since last year: retention

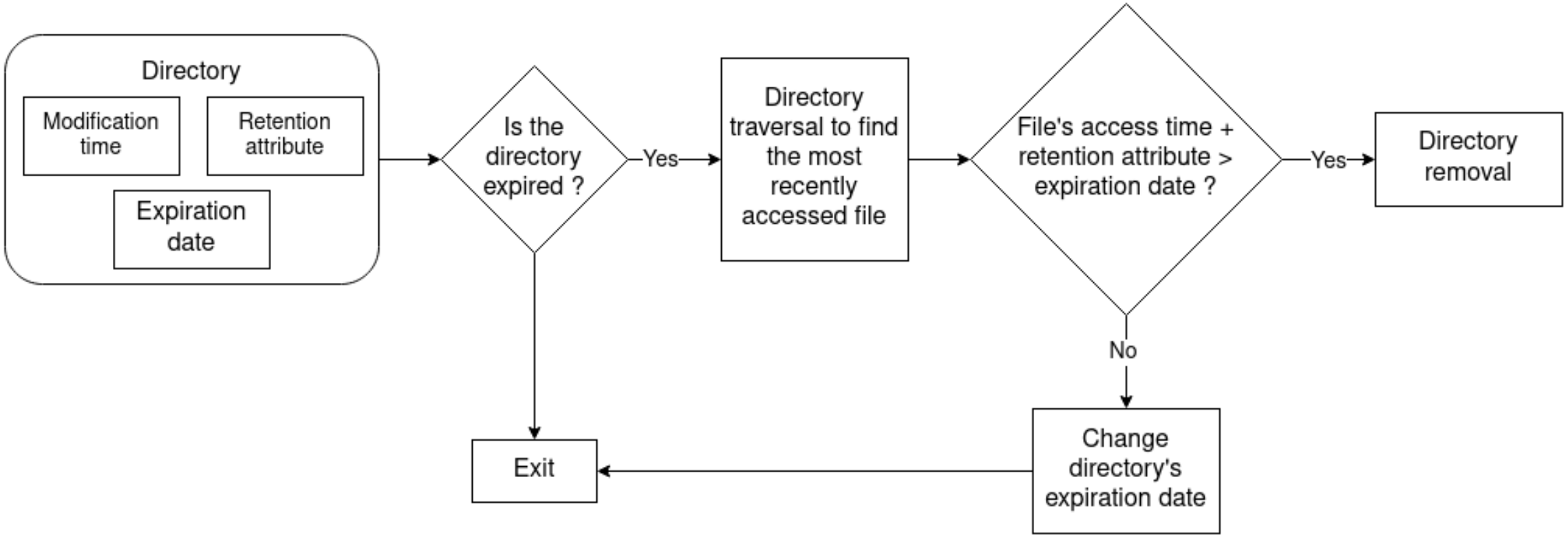
- Users do not want to archive or delete their old/unused files…
- Do it with an extended attribute set on their files
- **librobinhood**:
  - Improvements to the retention feature
    - Can be applied on a directory
    - User attribute used can be changed with the configuration

```
---
retention_xattr: "user.expires"
---
```

- **rbh-update-retention**:
  - Check if a directory is truly expired
  - If not, update its expiration date
  - Otherwise archive, delete, simple print, …

```
getfattr -d blob
# file: blob
user.expires="+42"
```

Retention on directory

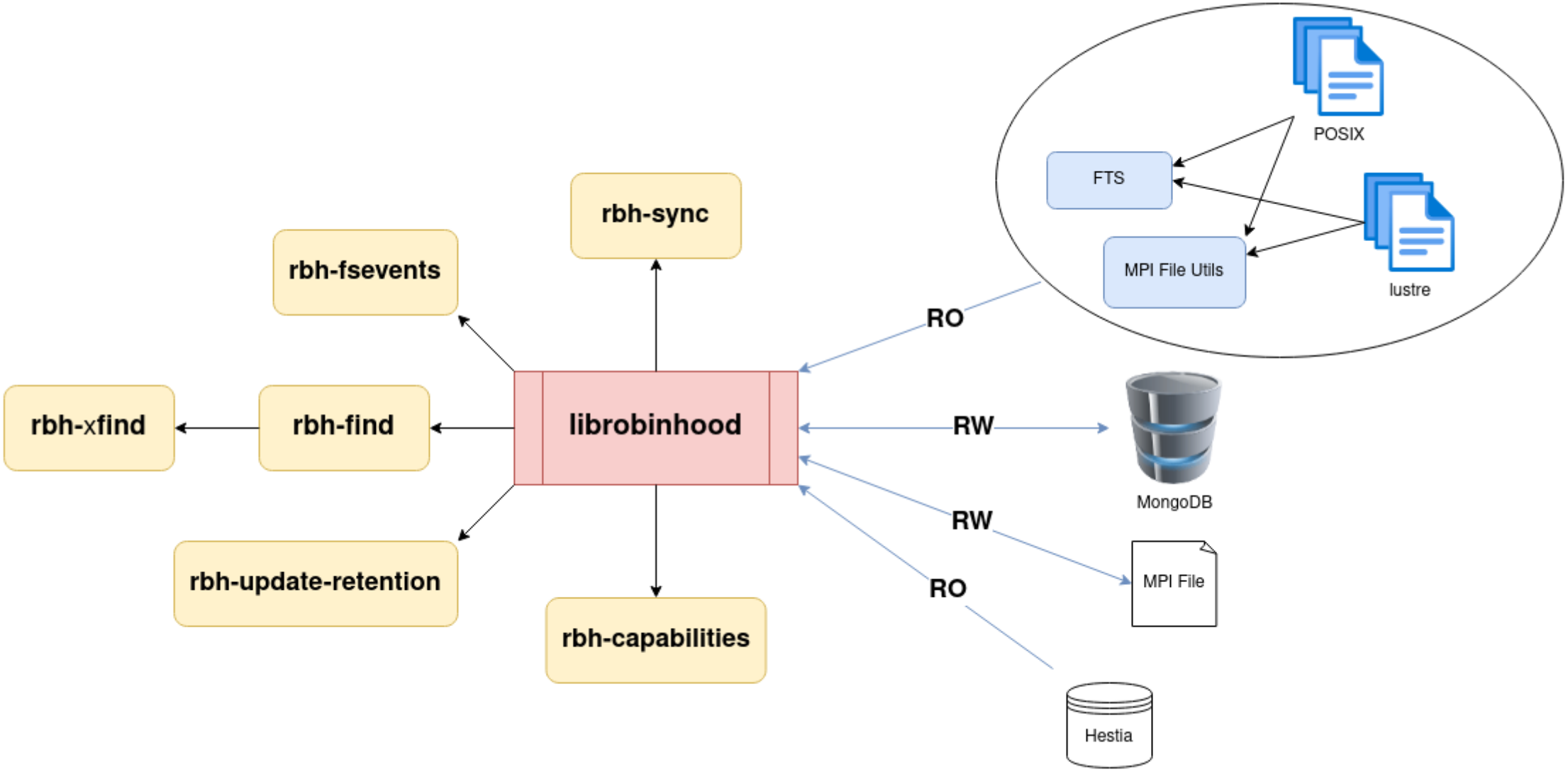


What's new since last year: retention

- **rbh-update-retention**:
  - Check if a directory is truly expired
  - If not, update its expiration date
  - Otherwise archive, delete, simple print, …

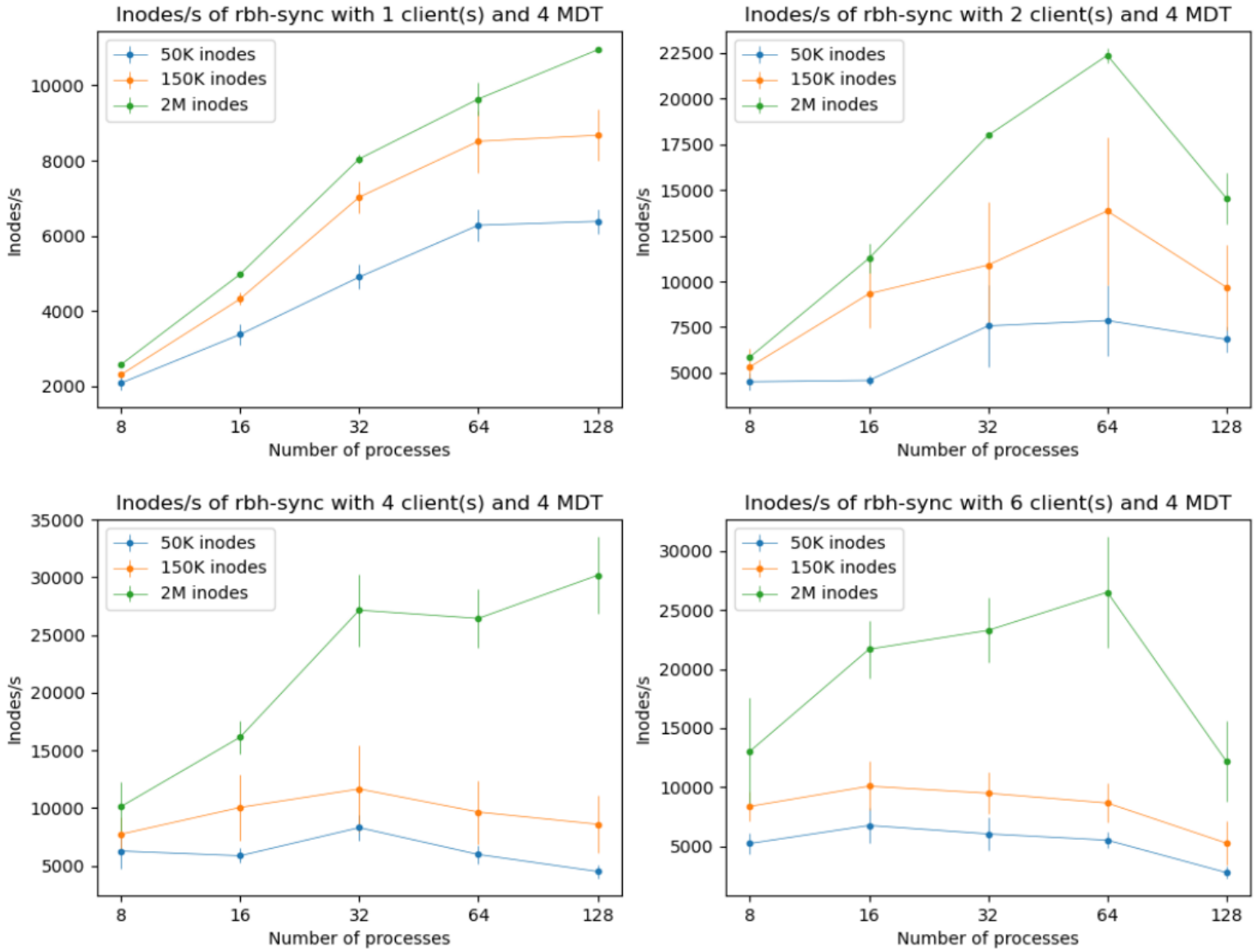
```
rbh-update-retention rbh:mongo:test_db /mnt/lustre

Directory '/dir1' expiration date is set to 'Fri Sep 20 08:26:48 2024'
The last accessed file in it was accessed on 'Fri Sep 20 08:26:49 UTC 2024'
Expiration of the directory should occur '+10' seconds after it's last usage
Changing the expiration date of '/dir1' to 'Fri Sep 20 08:27:00 UTC 2024'

Directory '/dir3' expiration date is set to 'Fri Sep 20 08:26:43 2024'
Directory '/dir3' has expired and it is empty, no other check needed
```

Current RobinHood architecture



Benchmarks

- Meet our performance requirements
- RobinHood v4 is scalable and works well in a multi-node and multi-client setup



Benchmarks

- 250 Millions inodes
- 4 MDT
- CPU:
  - Intel Xeon E5-2603 V3 1,6GHz 6 cores bi-socket
  - Intel Xeon CPU E5-2698 V3 2,30GHz 16 cores and 32 cores with hyperthreads bi-socket
- Memory:
  - 64GB RAM
  - 256GB RAM
- Memory: 64GB RAM and 256GB RAM
- Network: Mellanox MT27700 Family ConnectX-4 EDR ROCE V2
- Storage: 1 DDN SFA18KXE EDR ROCE V2
- Softwares:
  - Kernel 4.18.0-477.27.2.el8_8.x86_64
  - OpenMPI 4.1.1-3.el8
  - Lustre 2.15.5
  - MongoDB 4.2.3-3.0.1
  - RobinHood 3.1.7
  - MariaDB 5.5.68

RobinHood V3	RobinHood V4
286 GB	81.066 GB
1 node, 16 threads: 5 517 entries/sec	8 nodes, 12 processes per node: 6 293 entries/sec

In short

- MPIFileUtils backends
- Configuration file
- Retention feature
- **rbh-fsevents**

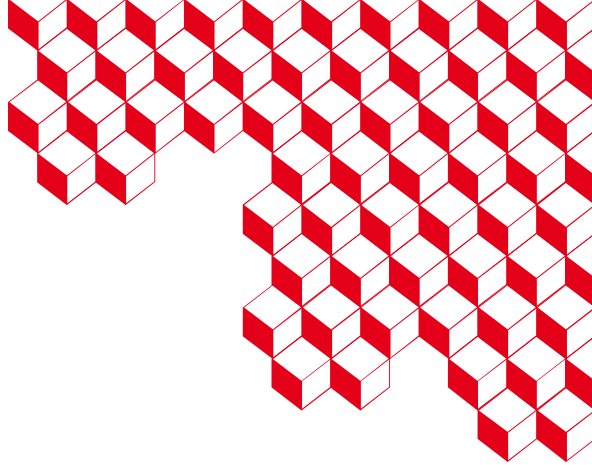
- RobinHood4 v1.0 is out!
- Pre-production tests done at CEA, specifically targeted for the retention feature → conclusive

What's next?

- **rbh-report**
- **Database improvements (sharding, configurable index, …)**
- **Performance improvements on the tools (multi-thread)**
- SQLite backend for usage by regular users
- Phobos backend for cold storage management
- SELinux backend
- Deployment of the tools on our systems during 2024/2025

Want to help with the development ?

- The tool suite is on Github: https://github.com/robinhood-suite/robinhood4
- Reviews done on Gerrithub
- Example of a patch in review:
  - https://review.gerrithub.io/c/robinhood-suite/robinhood4/+/1198697: patch to update the tests for newer versions of Alma Linux, OpenMPI and Lustres
- Feel free to install the suite and test it out for yourself, any feedback is appreciated!



Thanks for your attention

RobinHood v4 progress report

Yoann Valeri
Software stack developer at CEA
yoann.valeri@cea.fr

Why a version 4 for RobinHood?

	Version 3	Version 4
Scale up	SQL paradigm *MariaDB*	NoSQL paradigm *MongoDB*
Code genericity	Software specialised for Lustre filesystems	Generic tools calling specific backends
Inclusion to Linux repositories	Expert system	Library of features and applications
Code refactoring	Heavy code caused by Lustre behaviour evolution	Clean design to better correspond to current filesystems