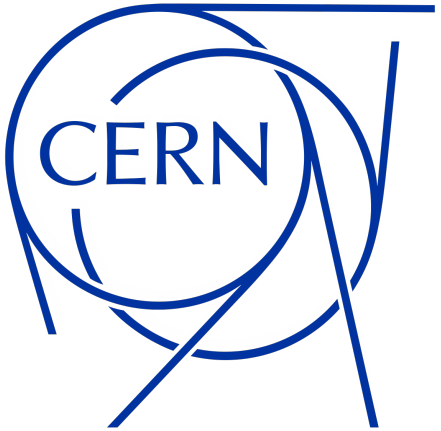


EOS 5.2 Status

Elvin Sindrilaru

on behalf of the EOS Team

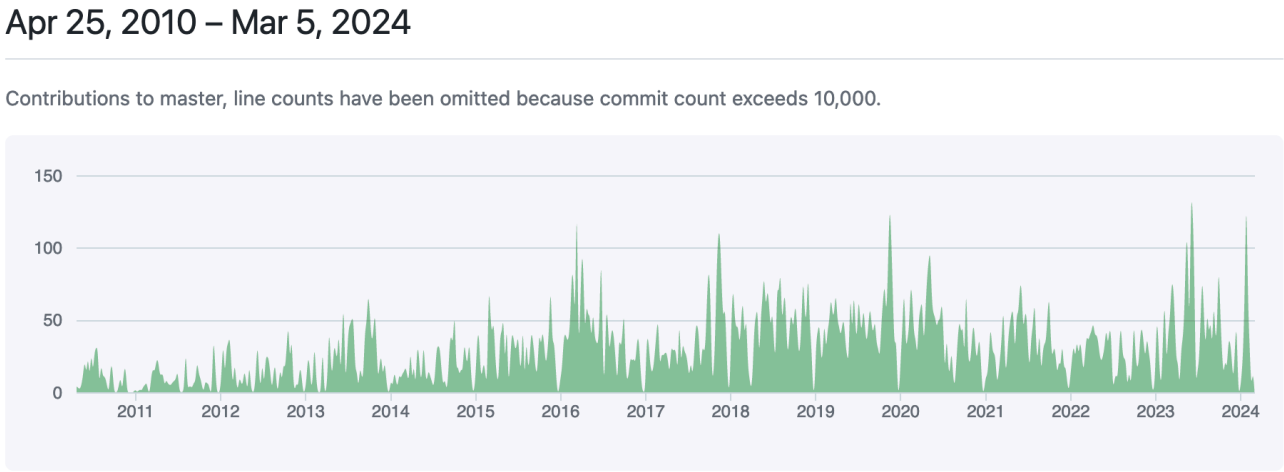
EOS Workshop - 15/03/2024

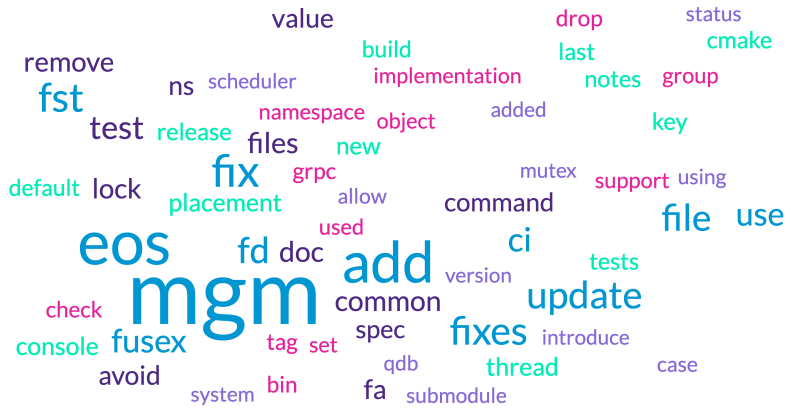
Outline

- Releases, supported OSes and packaging
- MGM API and command updates
- FST developments
- FUSE client
- Continuous integration and testing

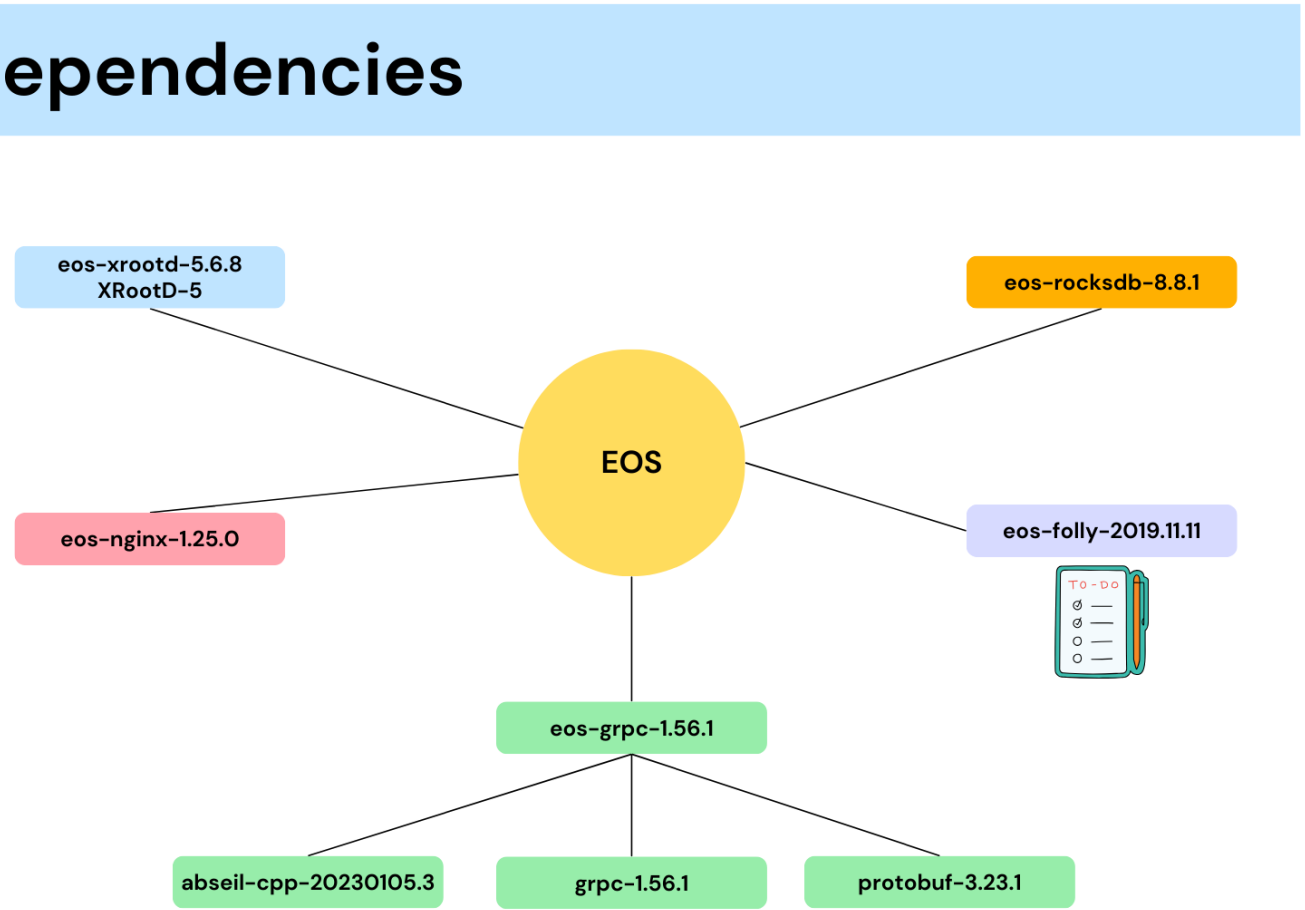
Releases and OS support

- **30 new releases - latest 5.2.19**
- **5.2 branch since October 2023**
- **Major dependencies updated**
- **Important deployment changes:**
  - Dropped **eosd** support
  - Deprecate **libmicrohttpd** use **XrdHttp**
  - Phase out use of **gridftp** and decommission **gateways**
- **EOS 5 deployed in all clusters at CERN**
  (physics instances + CTA + CernBox)
- **Operating Systems consolidation and support**
  - CentOS7 (**EOL June 2024**)
  - **Alma9** (> 50% migrated)





Update dependencies



MGM API and command updates

- **XRootD extended attribute support (5.0.0)**
  - Implemented in EOS >= 5.2.13
  - Consolidated internal EOS extended attributes API

- **Switch to new find command implementation**
  - Optimised for QuarkDB namespace implementation
  - Better handling of the MGM namespace cache
  - Old implementation still available: `eos oldfind`

```
$ xrdfs xattr <path> <code> <params>
    Operation on extended attributes. Codes:

    set    <attr>          Set extended attribute; <attr> is
                             string of form name=value
    get    <name>          Get extended attribute

    del    <name>          Delete extended attribute

    list                   List extended attributes
```

- **New commands and features**
  - `eos devices` - display storage device statistics based on S.M.A.R.T info
  - `eos ns benchmark` - run namespace metadata benchmark inside the MGM - *USE WITH CARE!*
  - `eos rclone` - tool for synchronising data between different EOS endpoints
  - `eos du` - display Linux-like `du` information for directories
  - `eos sched` - new scheduler configuration interface
  - **Glob functionality** for `rm` and `ls` commands

MGM internal changes

1 MGM data orchestration improvements

Incorporating feed back from Run3, data challenge exercises and regular operations

Groupbalancer

- extended functionality - free space engine
- add blocklist to exclude groups

FSCK extensions

- ubiquitous deployment
- new repair scenarios

File system balancer

- validate new design
- drop old components

2 Archive tool

Simple alternative to complex data management worflow with TAPE backend

Update internal API

- use new find API
- new xattr API

Forbid archival of files not supported by tape

- 0 size files
- version/atomic files

3 Third-party tools using EOS

Accomodate external requirements

WLCG tokens and Data Challenge

- token configuration
- deployment optimizations e.g. private key repeated downloads

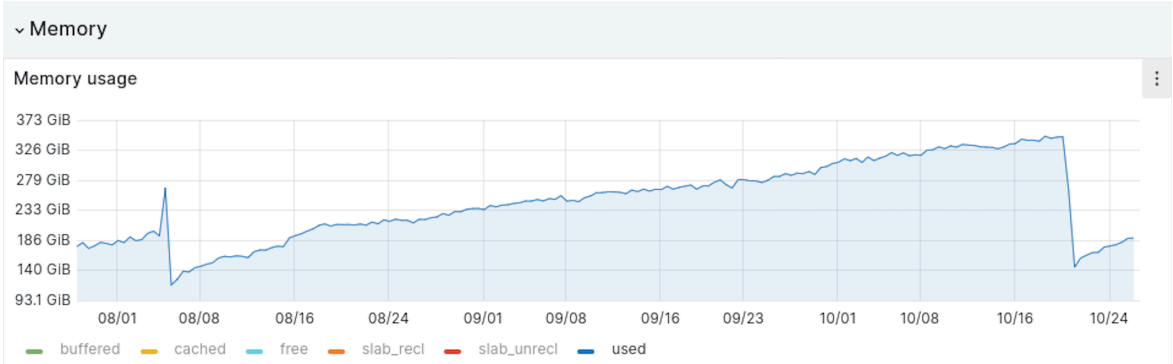
CERNBox deletion tracking for recyle bin

- sys.dtrace
- fix recycle bin cleanup bug

MGM main developments

- **New scheduler implementation**
  - **Different strategies**: round-robin/flat/weighted-round-robin

- **FileSystemView cache clean-up**
  - Keep the **memory overhead** under control
  - Used by **internal data orchestration engines**
    - drain, balance etc.



- **Namespace fine-grained locking**
  - Address **complex cases** such as `rename`
  - **Improve locking** for the `find` command
  - **Incremental improvements** for other ops in the pipeline

QuarkDB pub-sub messaging

- **Impacts all daemons/services in EOS**

- **MGM**
  - **GlobalConfigChangeListener** - avoid busy-loop implementation and use channel subscription and notification
  - **QdbErrorReportListener** - collecting error messages from the FSTs
  - Add **local hash specialisation** for groups and spaces that don't need to communicate with FSTs
  - **Removed unused functionality**: TransferJob/TransferQueue/TransferEngine/TransferMultiplexer, old FS balancer
  - **FileSystem and MessagingRealm** refactoring to register listeners and interests

- **FST**
  - Reduce chatter by using **batch updates** instead of individual messages for statistics
  - Adapt FileSystem registration/booting and updates coming via **QClient callbacks**

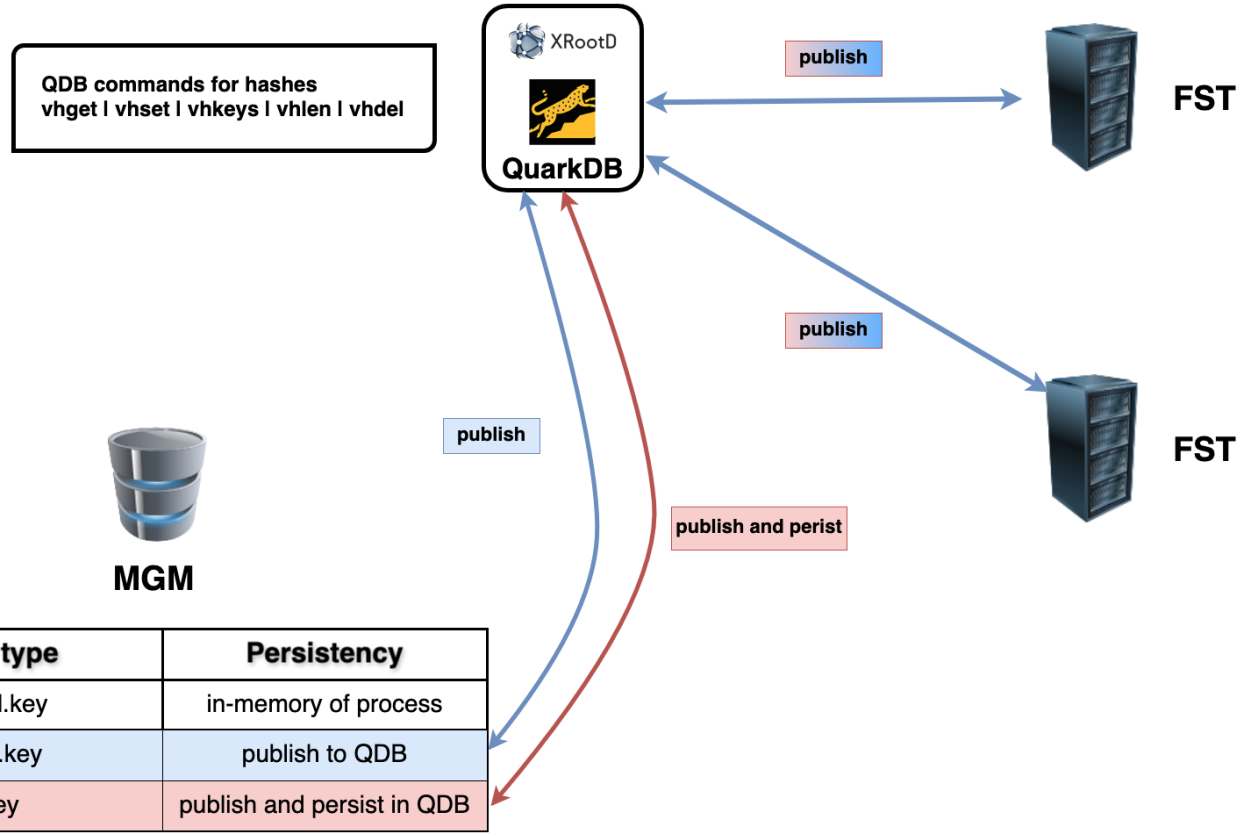
- **QuarkDB**
  - Update used API for **rocksdb 8.8.1**
  - Fix bugs related to **pub-sub and notification functionality**

Shared hash implementation

Shared hash characteristics

DATA STRUCTURES	LOCAL UPDATES	TRANSIENT UPDATES	PERSISTENT UPDATES
File System	✓	✓	✓
Node	✓	✓	✓
Group	✓		
Space	✓		

Message flow for shared hashes



Key type	Persistency
local.key	in-memory of process
stat.key	publish to QDB
key	publish and persist in QDB

- Built on top of a **PUB/SUB API** that can be used standalone e.g **error report listener**

FST developments

- **LevelDB support removed starting with eos-5.2.0**

- **Publishing info enhancements**
  - **RX/TX errors and dropped packet** counters
  - Avoid reliance on external shell scripts and use `syscalls` or `/proc/` info
  - **Publish all S.M.A.R.T. attributes** as compressed **JSON** object

- **FSCK extension for Erasure Coded (RAIN) files** - full file checksum corruption detection

- **HTTP chunk uploading fixes (XRootD)** to avoid infinite loop when client sends no data

- **Packet marking support (SciTags) -** via XRootD support

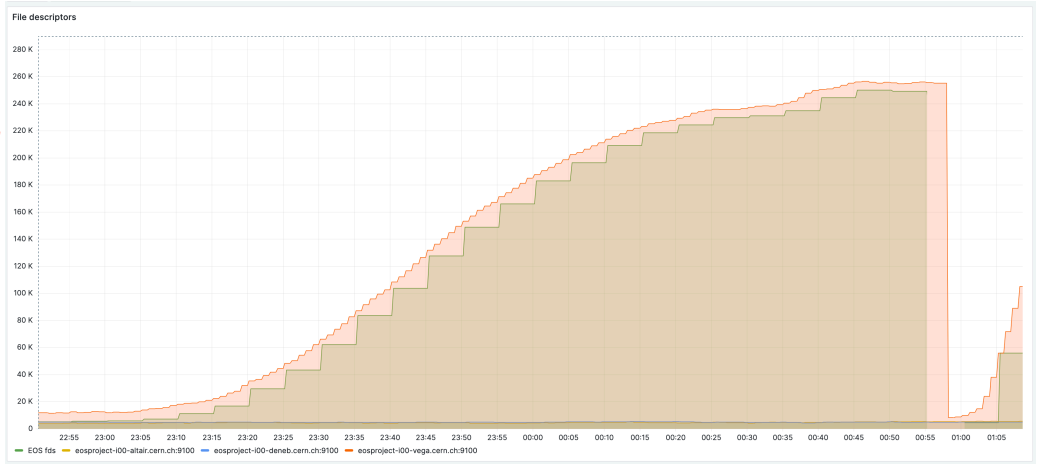
- **eoscp** tool enhancements
  - Add **checksum comparison** between source and destination
  - Add **JSON** output formatting

FUSE client

- **FUSE internal protocol improvements (v5)**
  - Less “**chatty**” protocol
  - Drop **BroadcastRelease** for deletions -> reduce load on the MGM
  - Full support for **ZTN** and **EOS tokens** through eosxd

- **Improve FSCK stats** by dropping know broken files
  - **Skip deletion through recycle bin** for files still open for writing

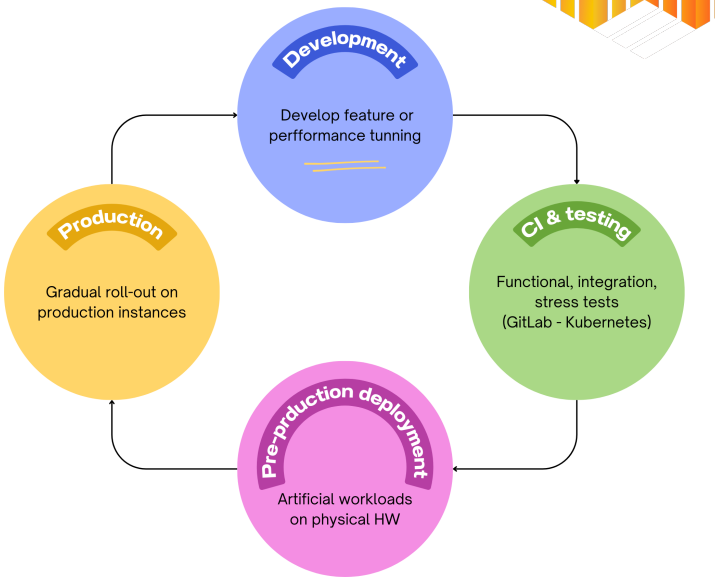
- **Important bug fixes**
  - Fix too **trivial hash function for credential cache**
    - Leading to **false sharing and TCP connection storm** towards the MGM
  - Locking and **concurrent access/update/read of string** objects
  - Avoid **interference between concurrent mount attempts** for same FS (Alma9)

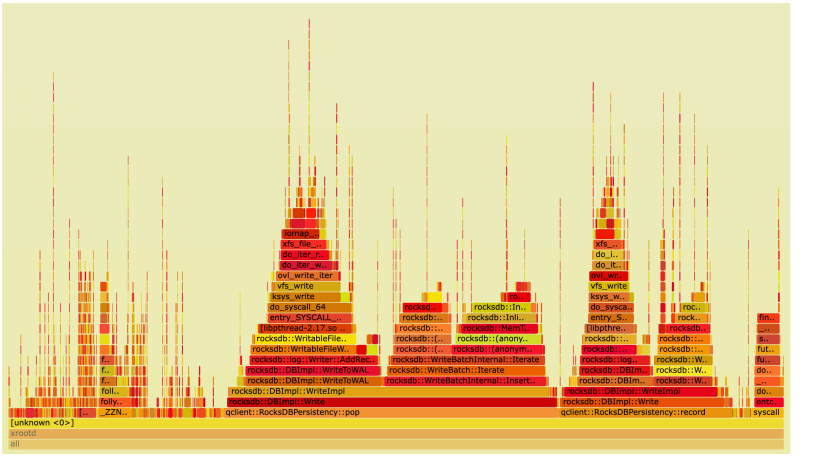


- Dedicated talk to the [evolution of eosxd]

Continuous integration and testing

- **Extended the HTTP test coverage**
  - HTTP TPC pull/push test with macaroons
  - Vanilla HTTP transfers with X509/tokens
  - EOS token tests for HTTP
- **Flamegraph charts from CI pipeline**
- **New CI job for ASAN and TSAN enabled packages**
- **New ARM builds for server + client**





Miscellaneous

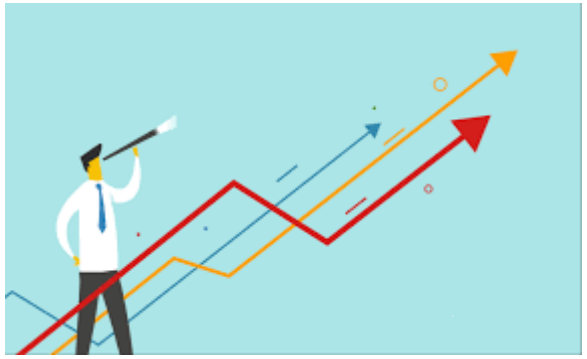
- **Extended documentation of file obfuscation and encryption**

- **Changes in behaviour when new node is added to the cluster**
  - Need **explicit MGM registration:** `eos node set eos-new-node.cern.ch:1095 on`

- **Document the need for quota node on the conversion directory**
  - `/eos/instance/proc/conversion`
  - If quota enabled this will be **enforced** at the closest quota node
  - Define a project quota for the conversion directory i.e **uid=99 gid=99**

- HTTP REST API **-** based on GRPC gateway
  - **OpenAPI** specification

Thank you! Questions? Comments?



home.cern