
CERN

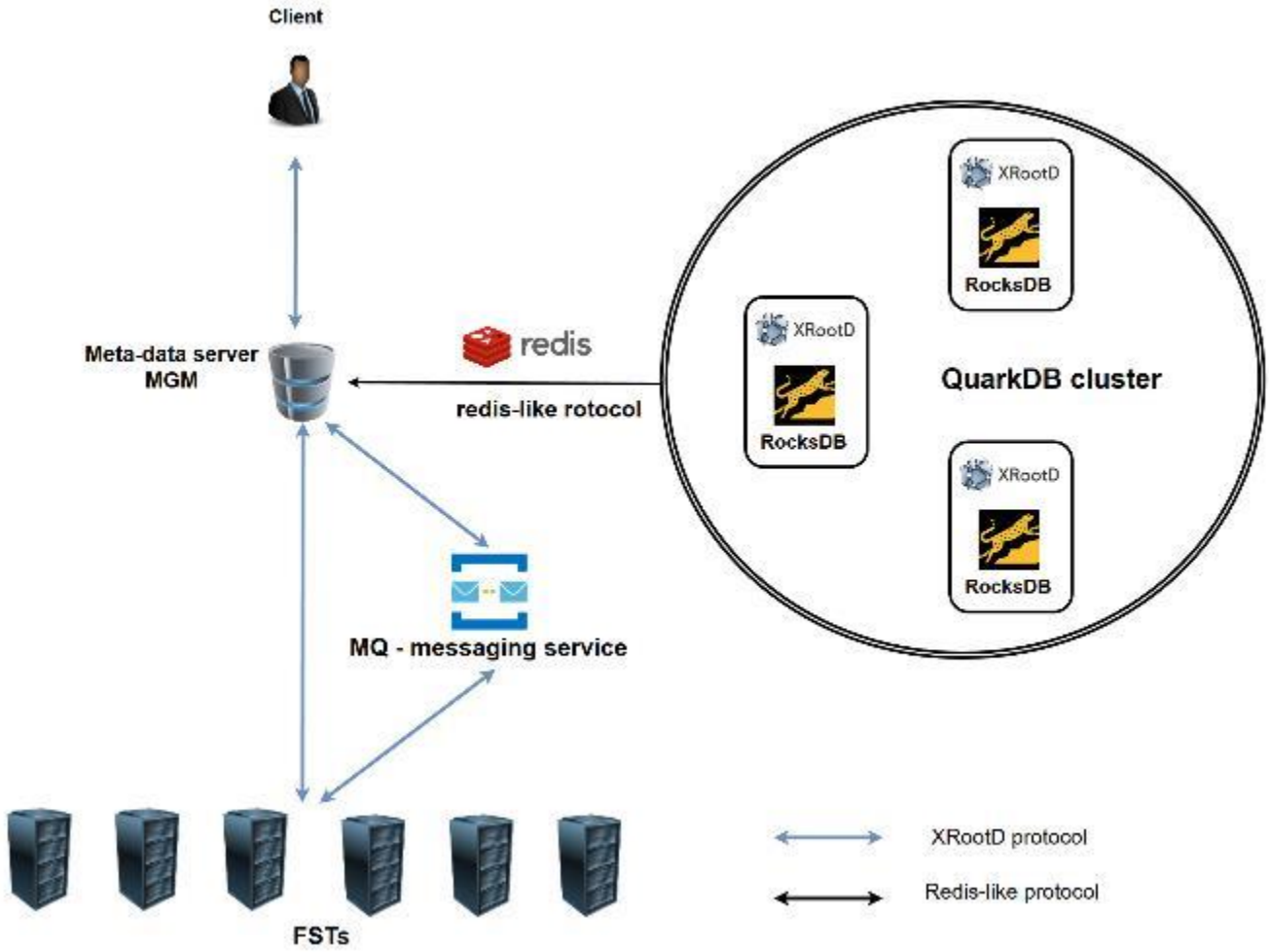
EOS Citrine updates and developments

Elvin Sindrilaru
on behalf of the **EOS team**

Outline

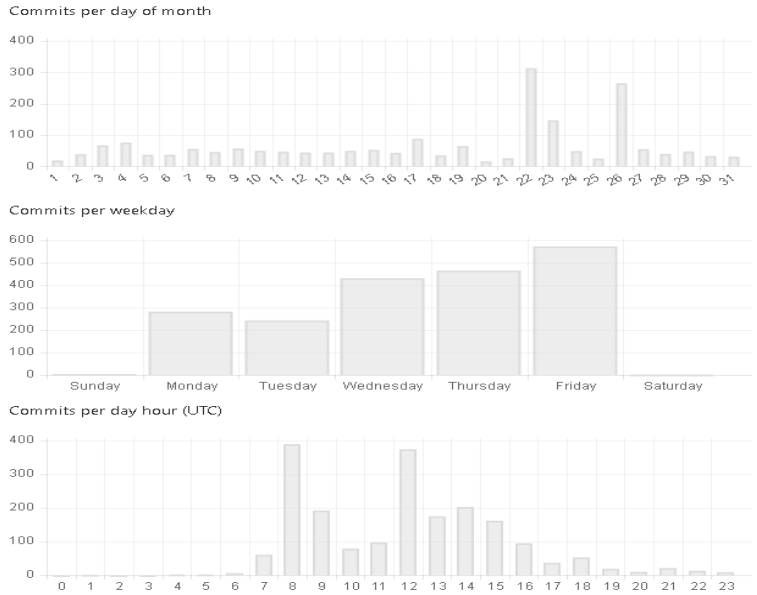
- EOS architecture overview
- New namespace and FUSEx
- Central draining
- Recycle-bin structure changes
- Console commands and Protobuf migration
- QuarkDB configuration and HA setup
- Plans for the future

EOS architecture



EOS releases and repositories

- Since the last workshop
  - **87 tags**
  - **~2300 commits**
- Latest releases
  - **Testing: 4.4.23**
  - **Stable: 4.4.18**



- **New testing repo**
  - http://storage-ci.web.cern.ch/storage-ci/eos/citrine/tag/testing/
- **Stable repo**
  - http://storage-ci.web.cern.ch/storage-ci/eos/citrine/tag/
  - Receives packages from the testing repo after running for **a few weeks in production at CERN**

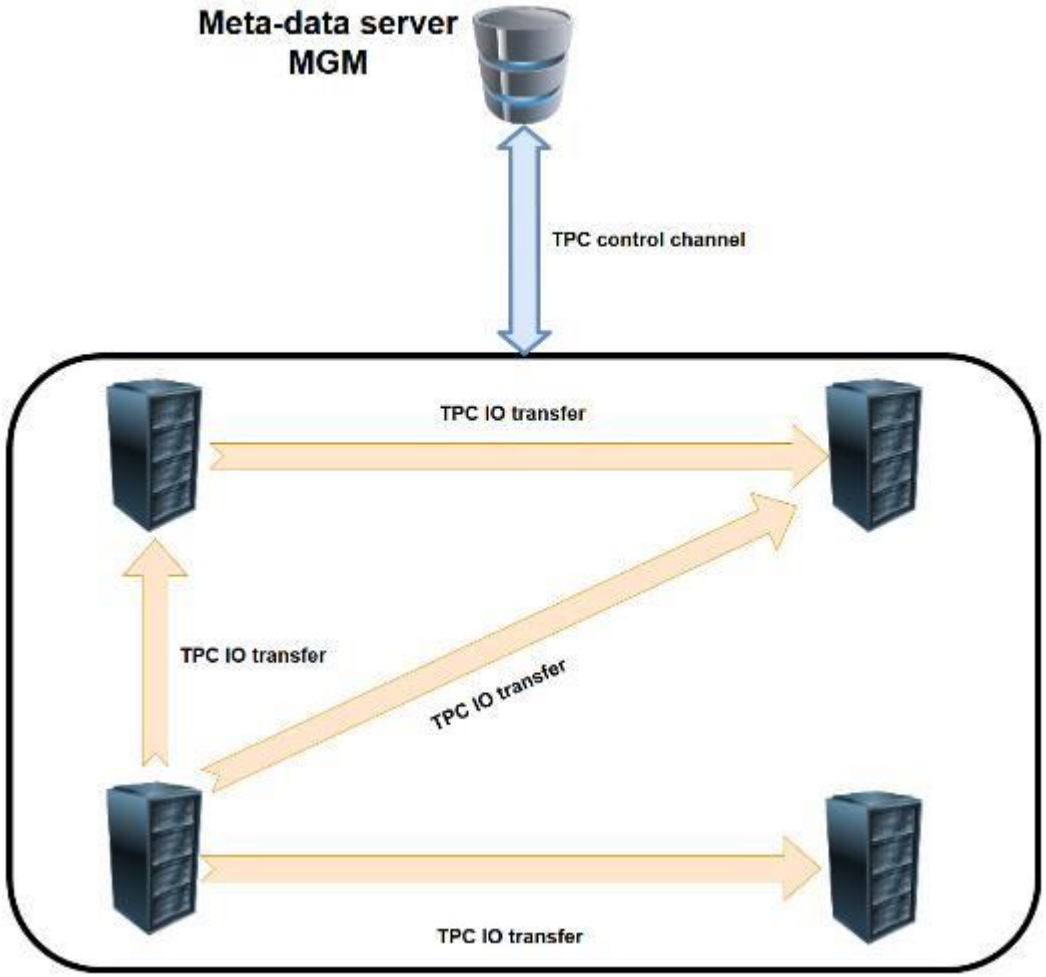
QuarkDB namespace and FUSEx

- **Namespace**
  - **QuarkDB in production** at CERN:
    - **EOSHOME** instance acting as backend for CERNBOX
    - **EOSBACKUP** holding > **1.5 B** files
    - **EOSPPS** holding > **3.5 B** files
    - Other experiment instances soon to follow …
  - Focus on **performance optimization**
    - Aggressive **prefetching** from the QuarkDB backend
    - Avoid locking during network requests
    - More details in Georgios's talk
- **FUSEx (eosxd)**
  - Ironing out corner-cases and subtle bugs
  - Again **focus on performance** – both client and server side
  - More details in Andreas's talk

Central draining

- **Old distributed draining** model not scalable for the new namespace
  - Each FST was querying repeatedly the namespace for the list of files to be drained
- **Central draining** now steers transfer from the MGM using **XRootD TPC transfers**
  - Simplify the code on the FST side
  - Automatic retries and fallback to other replicas if first attempt failed
  - Handles any type of layout: plain, replica, RAIN
  - Dedicated/configurable pool of threads doing the draining
  - Queue for pending file-systems to be drained
- In a better position to move out from the MGM code and create a **draining micro-service** … in the future

Central draining overview



Central draining configuration

- Dynamically configurable drain thread pool

  eos ns max_drain_threads <num>

- Other configuration saved as space attributes
  - **drainer.node.fs** – max number of file-system in draining per node
  - **drainer.fs.ntx** – max number of parallel transfers per file system
  - **drainer.retries** – max number of retries if failed transfers

- **Monitor** performance:

```
[root@eosbackup-ns-00 (mgm:master mq:master) ~]$ eos ns stat | grep Central
 all DrainCentralFailed          678.79 K     7.00    58.29    11.54     8.38     -NA-     -NA-
 all DrainCentralStarted          18.53 M    47.75   118.61    50.85    35.08     -NA-     -NA-
 all DrainCentralSuccessful       17.83 M    40.75    16.66    21.32    25.21     -NA-     -NA-
```

Recycle bin structure changes

- Existing recycle path convention:

/…/proc/recycle/gid/uid/dir1#:#dir2#:#file1.dat.hex_fid

- **Drawbacks:**
  - Flattens the entire recycle history for a user
  - Leads to extremely large directories (100k – 1M)
  - Considerable scalability issues when using the QuarkDB namespace
- **New recycle path convention**

/…/proc/recycle/uid:<val>/<year>/<month>/<day>/<hash>/path.hex_fid

Console commands and Protobuf integration

- Why change?
  - Problem **encoding** paths with “funny” characters
  - **“Snowball effect”** of request when XRootD client timeout triggered (60 seconds)
  - No way to **rate limit** the requests and avoid **starvation**
  - **Double parsing** of the same info on the console and server side

- Majority of **admin commands** moved to the new format

- **Old format:** `mgm.cmd=fileinfo&mgm.path=/eos/cms/path`

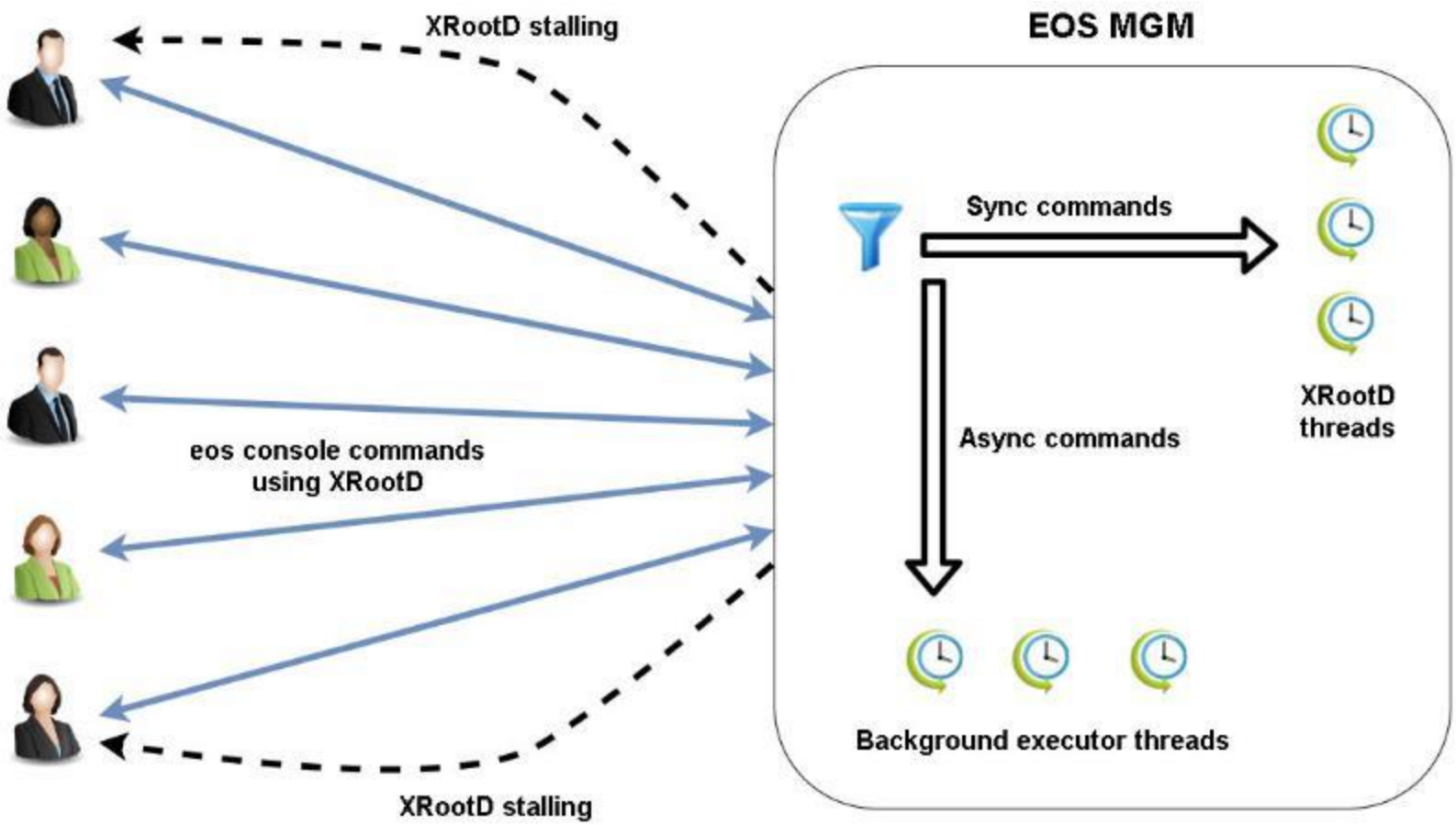
- **New format:** `mgm.cmd.proto= MgJSAA==`

Console commands features

- Use **Google Protobuf** messages
  - Ensures forward and backward compatibility
  - Base64 encoding of the serialized Protobuf message

- **Dedicated thread pool** on the MGM side serving only console commands
- **Decouple** the XRootD threads from **long running background operations** e.g. find, dumpmd etc.
- **Tracking** of client requests and detecting **request resubmission**
- **Queuing** of requests based on their type to avoid starvation of other clients

Console commands workflow



EOS configuration in QuarkDB

- Necessary step in providing **high-availability** setup
  - Move file-based config (default.eoscf) to QDB

- MGM setup requirements (**xrd.cf.mgm**):
  - **mgmofs.cfgtype quarkdb**
  - **mgmofs.qdbcluster <qdb1> <qdb2> …**
  - **mgmofs.qdbpassword_file <some/file>**
- **Configuration export** done using:

```
eos config export <path_to_config_file>
```

- Inspect the configuration directly from QuarkDB

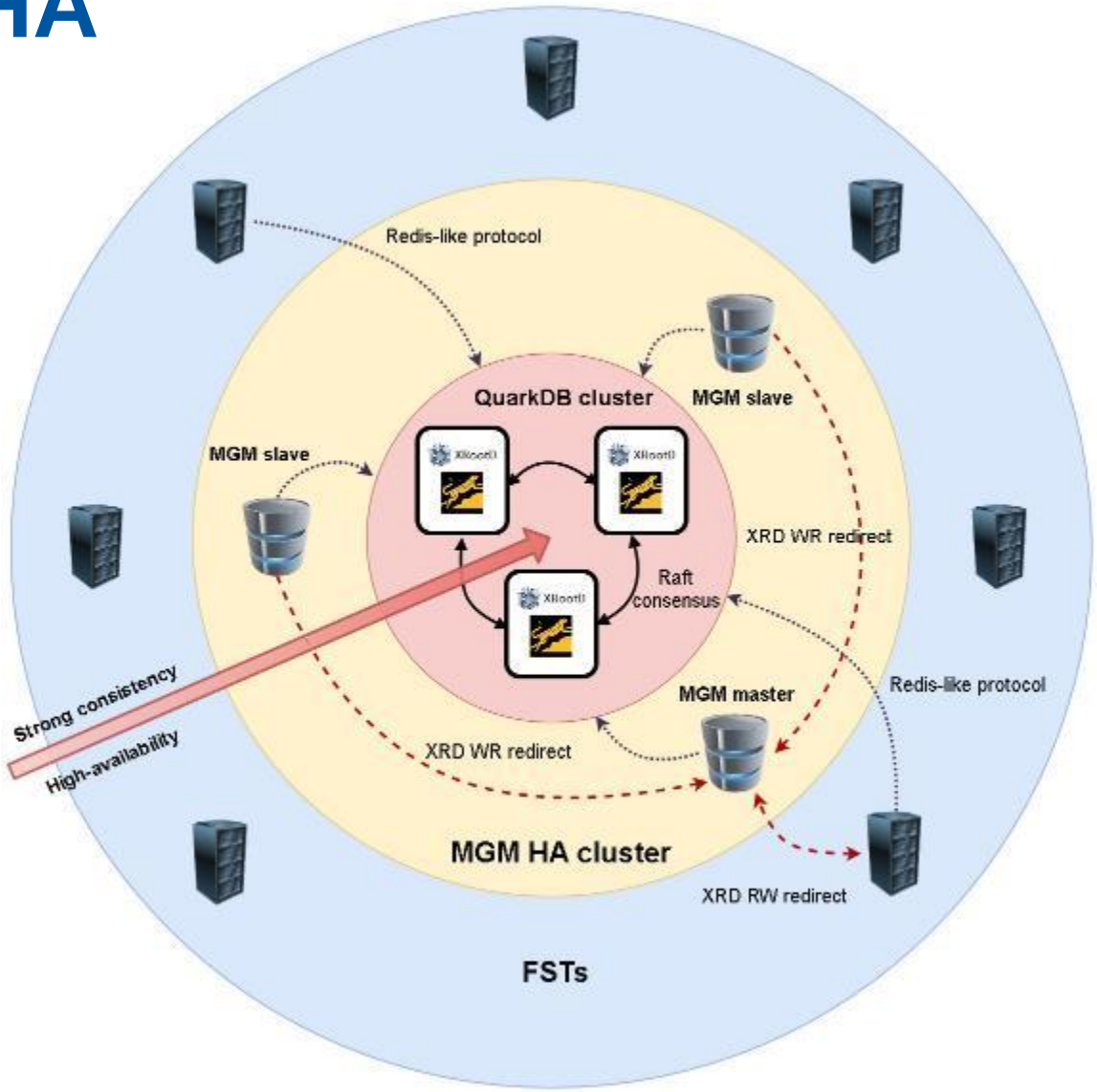
```
[root@eospps-ssd-ns1 ~]# redis-cli -p 7777 keys "eos-config:*"
1) "eos-config:backup1"
2) "eos-config:default"
[root@eospps-ssd-ns1 ~]# redis-cli -p 7777 hgetall "eos-config:default" | head -n 2
fs:/eos/lxfsre03a02.cern.ch:1095/fst/data01
```

QuarkDB leases

- Building block for providing **HA** for the MGMs
- Stores information concerning:
  - **Current owner** of the lease
  - **Validity** of the lease
- Operations on leases:
  - **lease_acquire**
  - **lease_release**
  - **lease_get** -> display information about the lease
- Master-slave MGMs synchronize using the lease key "**master_lease**"

```
[root@eospps-ssd-ns1 ~]# redis-cli -p 7777 lease-get "master_lease"
1) HOLDER: eospps-fe1.cern.ch:1094
2) REMAINING: 9023 ms
```

EOS HA


Redis-like protocol
QuarkDB cluster
MGM slave
MGM slave
XRD WR redirect
Raft consensus
Strong consistency
High-availability
MGM master
Redis-like protocol
XRD WR redirect
MGM HA cluster
XRD RW redirect
FSTs

EOS master-slave HA

- Rely on the **QuarkDB** lease to decide who is the master
  - Lease is **valid for 10 seconds**
  - Master **renews** the lease every **5 seconds**
- During a slave->master transition **reload the configuration from QuarkDB**
- Automatically **enforce/disable stall rules**
- Force a master to abandon the lease

```
eos ns master other
```

- Master-slave info displayed in the “ns” command
  - **ALL Replication is_master=true master_id=eos-mgm-1.cern.ch:1094**
  - **ALL Replication is_master=false master_id=eos-mgm-1.cern.ch:1094**

Plans for the future

- Stop support for the **beryl_aquamarine** branch: **June 2019**

- Focus on **stability** and better **fault-tolerance**
- Improve **availability** and **self-healing mechanisms**
  - Redesign the FSCK functionality

- **No big changes** from the current model

Plans for the future

- Pick a name for the next major release?

Danburite

Demantoid

Dravite

www.cern.ch