EOS 5 developments

Elvin Sindrilaru

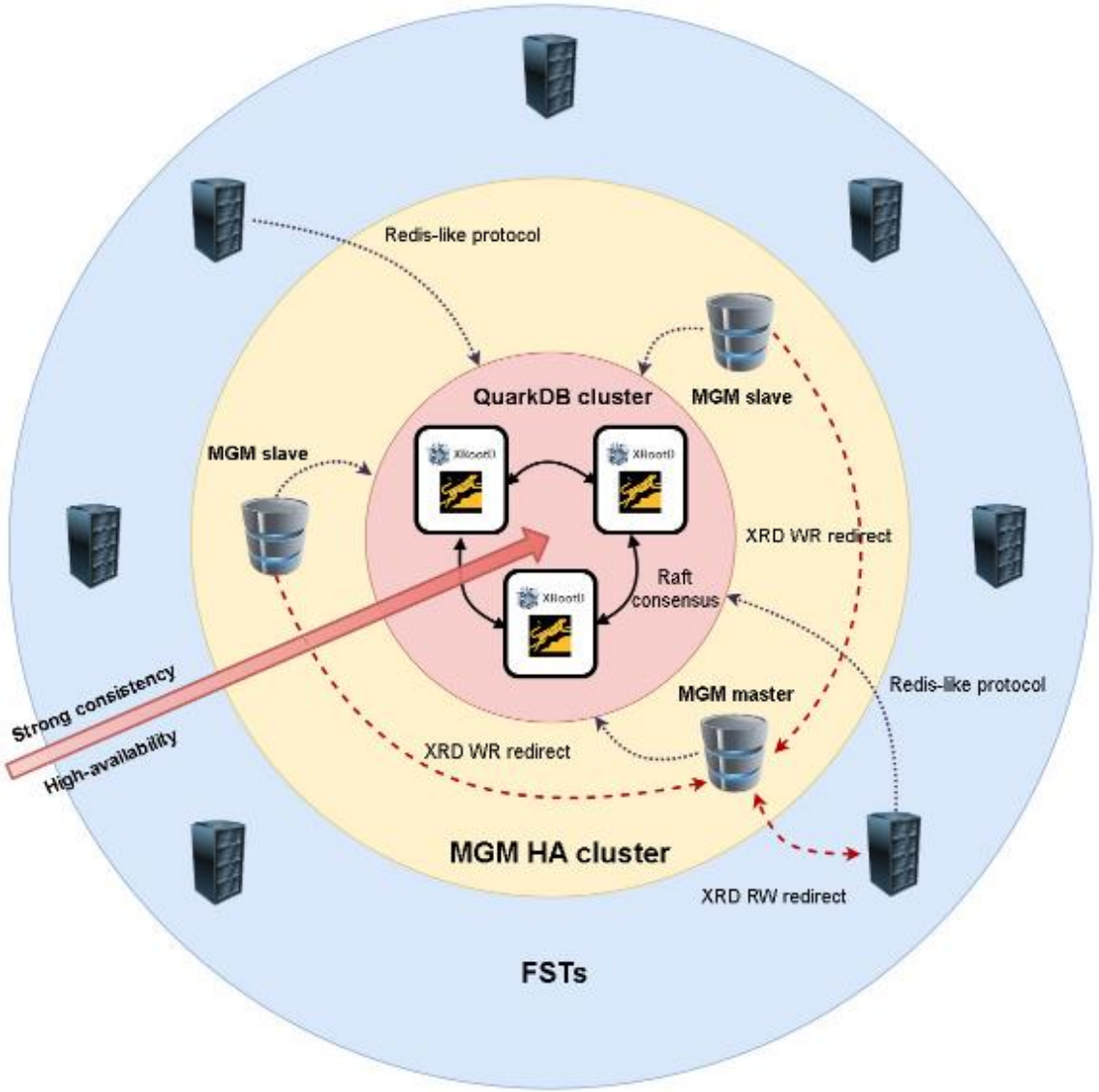
on behalf of the **EOS team**

25.04.2023

Outline

- General considerations
- MGM developments and new functionality
- FST developments and new functionality
- Client/FUSE developments

EOS architecture



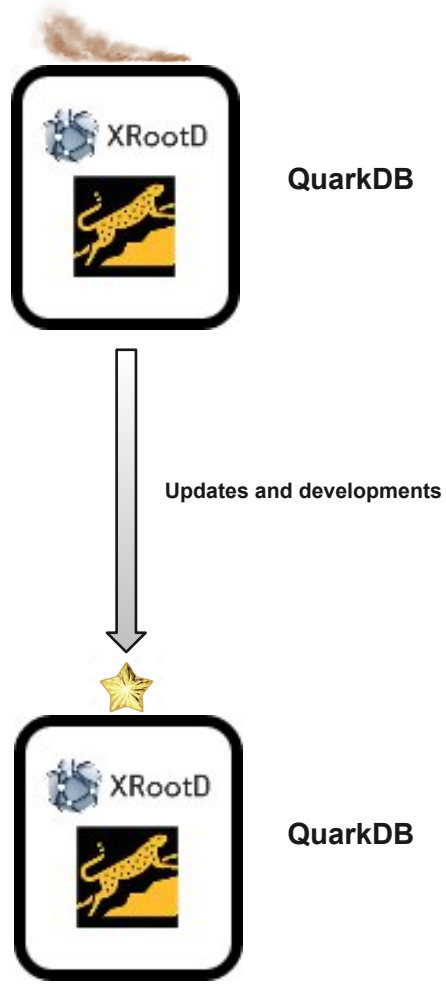
General considerations

- **Builds for new OSes**
  - Added builds for **Alma8/9, RH8/9, Fedora 36/37**
  - Added builds for **Ubuntu Jammy 22.04**
- **Move from Python2 to Python3**
  - Hit some nasty bugs in the **XRootD Python bindings** - fixed
  - Affected the **eos archive tool** and other helper scripts
- **Continuous improvement of the testing infrastructure**
  - **Drop CPPUnit** and rely only on **gtest**
  - Add end-to-end tests for **HTTP/Tokens/TPC**
  - Updated the **HELM deployments** to run all CI tests
- **Focus**
  - How to **best handle overloads** at the MGM level - stalling/delay?
  - How to **prioritise data-taking workflows** over other activities
  - **Analysis, optimization and tackling scalability/bottleneck** issues

Analysis and optimization

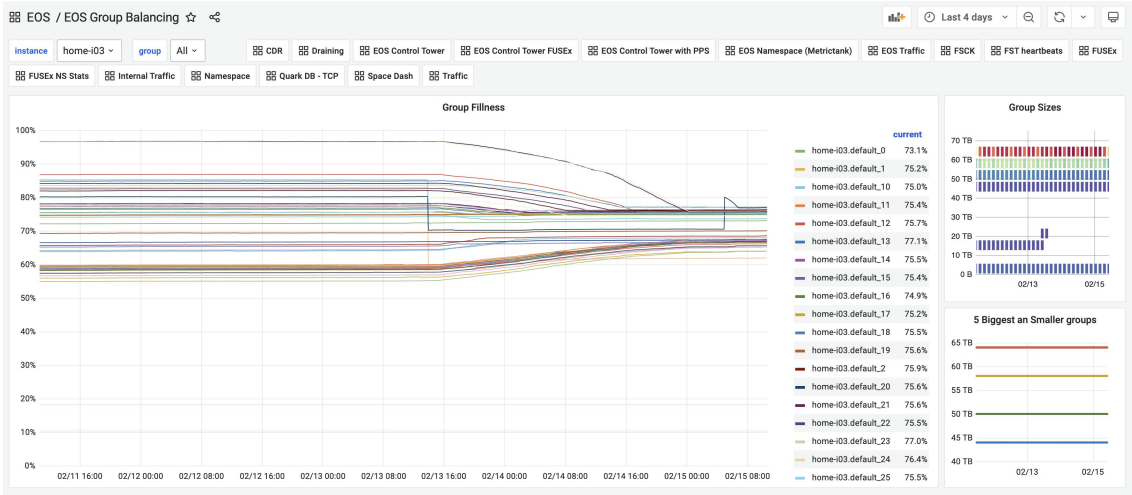
QuarkDB updates

- **Release and packaging**
  - **QuarkDB** now distributed as part of the **EOS** releases
  - New package: `eos-quarkdb-<version>.rpm`
  - **No strong requirement to update** the QuarkDB daemon unless specified in the release notes
  - Small fixes for **Alma 8/9 releases**
- **Developments**
  - No major development effort, small fixes, runs **very stable** in production
  - **Fix data race** in the publish-subscribe mechanism for shared objects
  - Allow running commands with **more arguments from the localhost** (i.e. unauthenticated)
    - e.g `redis-cli -p 7777 sscan fsck:orphans 0 COUNT 100`
- **Plans for the future**
  - **Update** the underlying **rocksdb** package version – currently running on 6.2.4 (Sep. 2019!)
    - Hit **sst files not being cleaned up bug** due to iterator objects pinning such files
  - **Update** the underlying **folly library** – currently running 2019.11.11
    - Hit **bug related to IOThreadExecutors** crashing when deleted



MGM developments (1)

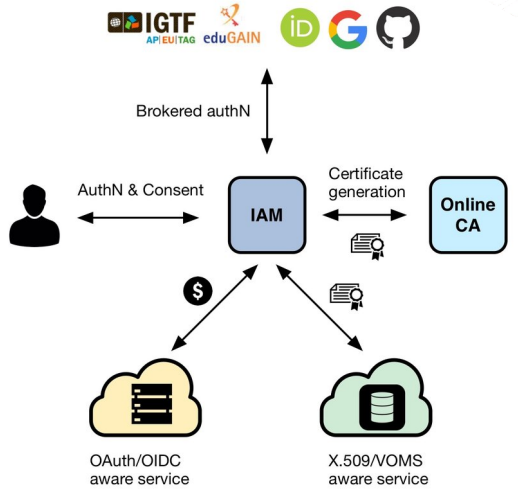
- **GroupDrain functionality**
  - **GroupBalancer** functionality extended and refactored
  - Handy when **retiring/shuffling disks/capacity**
  - Leverages the **Converter** functionality
  - Adds **Observer interface** for handling and dispatching **notifications**
  - **More details**: https://indico.cern.ch/event/1227241/contributions/5348018/



- **Central file system balancer (FsBalancer)**
  - Re-uses the internal **Third-Party-Copy job** for transfers
  - Allows to **drop the old bash-based TPC** implementation
  - Allows to **retire the TransferJob/Multiplexer/Queue** mechanism
- **Update quota** when converting to different layout – fix **quota accounting** issues (CMS)

MGM developments (2)

- **General token support**
  - Added support for **tokens over the xroot protocol (ztn)**
  - Added support for **eos tokens** over **HTTPS**
  - Refactored the **chaining of token authz** libraries
  - Drop the need for `eos-scitokens` package using default library from XRootD
  - **More details:** https://indico.cern.ch/event/1227241/contributions/5349673/
- **EOS token related functionality**
  - Allow **black/white-listing of token** vouchers
    - Implementing basic **revocation of tokens**
  - **tokensudo policy** - control when a token maps to the embedded uid/gid
    - **never** – only token permissions are taken into consideration
    - **strong** – not with unix protocol
    - **encrypted** – only with https, sss, ztn
    - **always** - no matter the protocol
  - **CLI command**: `eos vid tokensudo always|strong|encrypted|never`




```
// server issues a scoped token binding to a user/group
TOKEN=`eos token --path /eos/user/www/ --permissions rwx \
       --expires 16000000 --owner user1 --group group1
// export the token in the environment
export XrdSecsssENDORSEMENT=$TOKEN
// check the mapped identity
eos whoami
Virtual Identity: uid=1001 (99,3,1001) gid=5001 (99,4,5001) [authz:sss] \
                  sudo* host=localhost domain=localdomain
{
 "token": {
  "permission": "rwx",
  "expires": "1600000000",
  "owner": "user1",
  "group": "group1",
  "generation": "0",
  "path": "/eos/user/www/",
  "allowtree": true,
  "vtoken": "",
  "origins": []
 },
}
```

MGM developments (3)

- **IoStat functionality refactoring**
  - Moved collected statistics to **QuarkDB** – make **MGM stateless**
  - Monitor **time to completion** for transfers
  - Introduced New Year's bug that restarted all the MGMs
  - Added **caching** to reduce the load on the QuarkDB

```
┌─> Transfer (tf) sample info every 5 min: tf time for 90/95/99% of data, max tf and report times, average tf size, tf count.
```

io	application	90% [s]	95% [s]	99% [s]	max [s]	max report [s]	avg tf size	tf #	sample end time
out	eoscp	3	3	4	4	2	104.66 M	61	Tue Apr 26 11:25:18 2022
out	eos/gridftp	679	717	747	754	2	1.07 G	10	Tue Apr 26 11:24:17 2022
out	eos/converter	0	0	0	0	0	0	0	Tue Apr 26 11:23:49 2022
out	eos/replication	0	0	0	0	0	0	0	Tue Apr 26 11:24:49 2022
out	fuse	0	0	0	0	0	0	0	Tue Apr 26 11:23:15 2022
out	other	475	4.53 K	17.62 K	26.19 K	2	96.19 M	1.04 K	Tue Apr 26 11:22:51 2022
out	fuse::lxplus	0	0	0	0	0	0	0	Tue Apr 26 11:25:39 2022
out	fuse::bi	11	12	16	752	2	3.87 M	92.69 K	Tue Apr 26 11:22:34 2022
out	fuse::amssoc	23	29	38	44	2	1.54 K	125	Tue Apr 26 11:24:12 2022
out	tpc	0	0	0	0	0	0	0	Tue Apr 26 11:23:04 2022
in	eoscp	17	18	19	20	2	2.40 G	25	Tue Apr 26 11:23:09 2022
in	eos/gridftp	78	88	102	116	2	324.98 M	49	Tue Apr 26 11:23:07 2022
in	eos/converter	0	0	0	0	0	0	0	Tue Apr 26 11:23:54 2022
in	eos/replication	0	0	0	0	0	0	0	Tue Apr 26 11:24:49 2022
in	fuse	0	0	0	0	0	0	0	Tue Apr 26 11:25:13 2022
in	other	30	32	33	33	2	345.02 M	15	Tue Apr 26 11:23:06 2022
in	fuse::lxplus	15	16	16	16	1	365	1	Tue Apr 26 11:21:03 2022
in	fuse::bi	2	3	3	3	2	7.26 M	38	Tue Apr 26 11:23:11 2022
in	fuse::amssoc	30	35	40	41	2	299	52	Tue Apr 26 11:22:22 2022

- Fix **deadlock** triggered when **report files are rotated**
  - Race-condition triggered when writing to a closed file

- Finalizing the **Tape Rest API**
  - Implemented **bulk prepare requests**
  - Being **adopted** by other major storage providers
  - Released June 2022

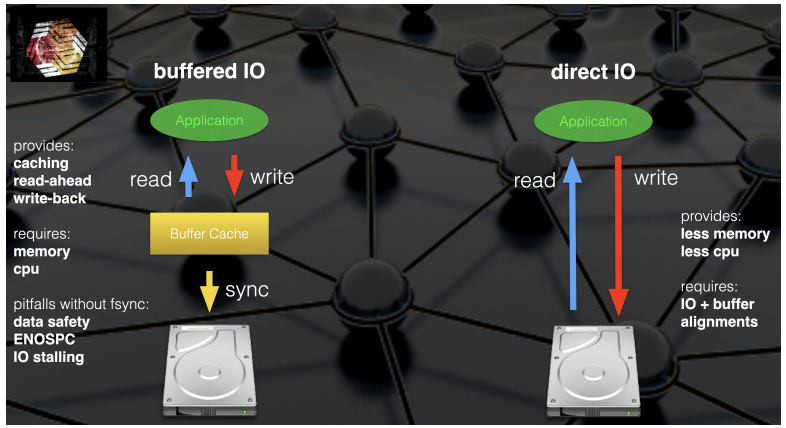
Tape Rest API

EOS

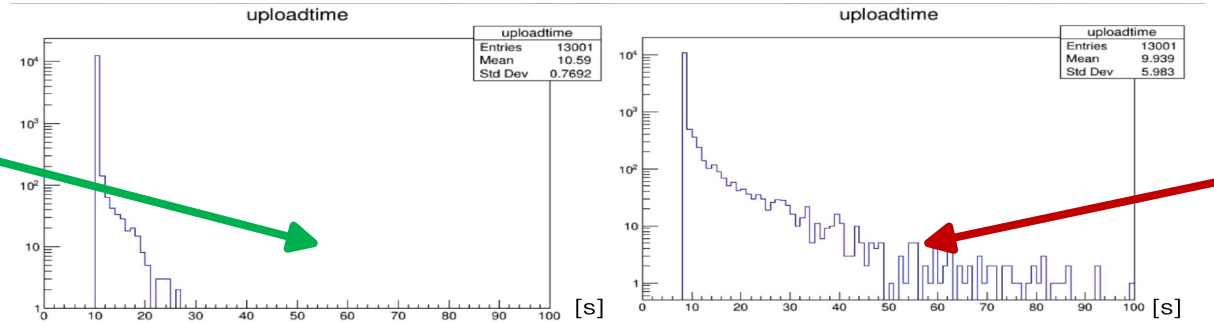
CERN Tape Archive

MGM developments (4)

- **Fine grained IO policies** defined by
  - **Space << Group << User << Application << Directory xattr**
  - IO types: **direct** vs. **sync** vs. **csync**
  - IO priorities: **real-time (rt) 0 >> rt 7 >> best-effort (be) 0 >> be 7 >> idle**
  - Default IO priority: **be 4;** Scanner IO priority: **be 7**
  - **Bandwidth regulation** – leads to reduced IO tails
  - **HEPiX 2022 talk**: https://indico.cern.ch/event/1123214/contributions/4809924/



With bandwidth regulation



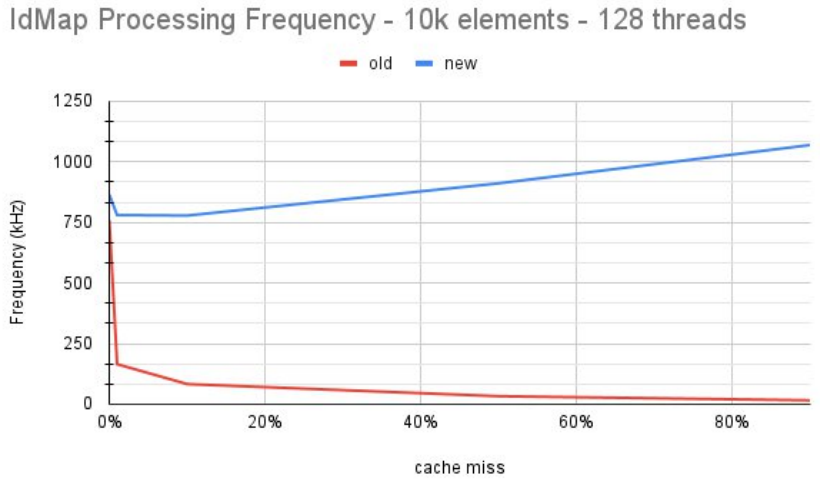
NO bandwidth regulation

- **File system scheduling overload**
  - **Avoid accumulation of streams** on slow file systems
  - Set **max num. of rd/rw streams** per file systems: `eos space config default space.max.wopen=200`
- **Meta-data overloads** mitigated by enforcing thread-pool limits per user

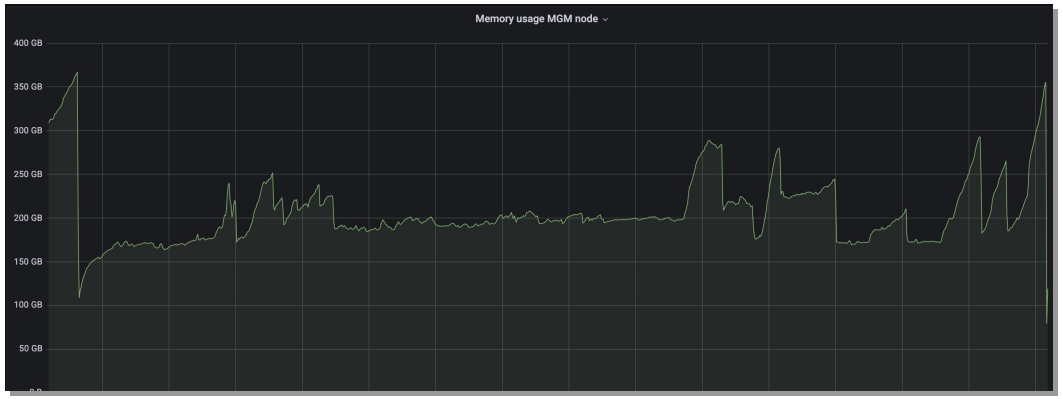
MGM developments (5)

- **IdMap refactoring and better caching**
  - Drop use of XrdOucHash and use **sharded caches**
  - Consolidated **token handling** for both **HTTPS** and **XRootD**
  - Add **KEYS mapping** via `x-gateway-authorization` header for **HTTPS**

```
eos vid set map -https key:211f82e6-beef-adda-dead-2449e4f1234 vuid:58602
```



- Fix **memory leak in the MGM** when using **regex data structures**
  - When printing and broadcasting **fusex capabilities**
  - Memory leak correlated with **user activity**



- Better enforcement of the **scatter policy**
  - Enforced via **environment variable**
    - `EOS_SCATTERED_PLACEMENT_MAX_ATTEMPTS`
  - Affecting **geo-distributed instance** (AARNet)

MGM developments (6)

- Add **atime** support for files
  - Set update **threshold as a space parameter**
  - Update atime once every 7 days

```
eos space config default space.atime=6048000
```

- Add new **fileinfo Status** information
  - Reflects the state of the file from the **fuse client perspective**
  - Helpful to spot **systemic issues or bugs**
  - Coupled with new file xattrs **sys.fuse.state** and **sys.fs.tracking**

```
eos fileinfo /eos/dev/opstest/esindril/file1.dat
  File: '/eos/dev/opstest/esindril/file1.dat'  Flags: 0644
  Size: 1490
Status: healthy
Modify: Sun Apr 23 19:15:48 2023 Timestamp: 1682270148.505499142
Change: Sun Apr 23 19:15:48 2023 Timestamp: 1682270148.505169673
Access: Sun Apr 23 19:15:48 2023 Timestamp: 1682270148.505400048
```

```
eos attr ls /eos/dev/opstest/esindril/file1.dat | grep tracking
sys.fs.tracking="+30937+29227"
```

Type	Description
locations::uncommited	File is being written to
locations::incomplete	Not all commits received
locations::overreplicated	More stripes/replicas then nominal layout
fuse::needsflush	Data still on the client side
fuse::reparing	Fuse repair process triggered
fuse::missingcommits	Fuse access and not all commits received

MGM developments (7)

- **QClient RTT and peak measurements**
  - Helpful to spot **QuarkDB or network issues**
  - **Real-time** indicator for the **health** of the instance
- Fix **crash in ContainerMD files/sub-dirs iterator** being invalidated
  - Triggered by **concurrent additions/removals** of entries to/from a directory
  - Counting the **number of buckets in a dense-hash-map** is not enough
- Take into consideration the **XRD_APPNAME**
  - Might have back-fired on us since now output is verbose
- **Newfind command** performance and scalability improvements
  - Will soon replace the current **eos find**
  - Can **by-pass the LRU cache** in-memory – avoid **cache trashing**
- Add **secondary group access** control
  - Env. variable `EOS_SECONDARY_GROUPS=1`

```
# QClient RTT should always be very low
eos ns | grep QClient
ALL      QClient overall RTT              0ms (min)  0ms (avg)  5687ms (max)
ALL      QClient recent peak RTT          0ms (1 min) 0ms (2 min) 0ms (5 min)
```

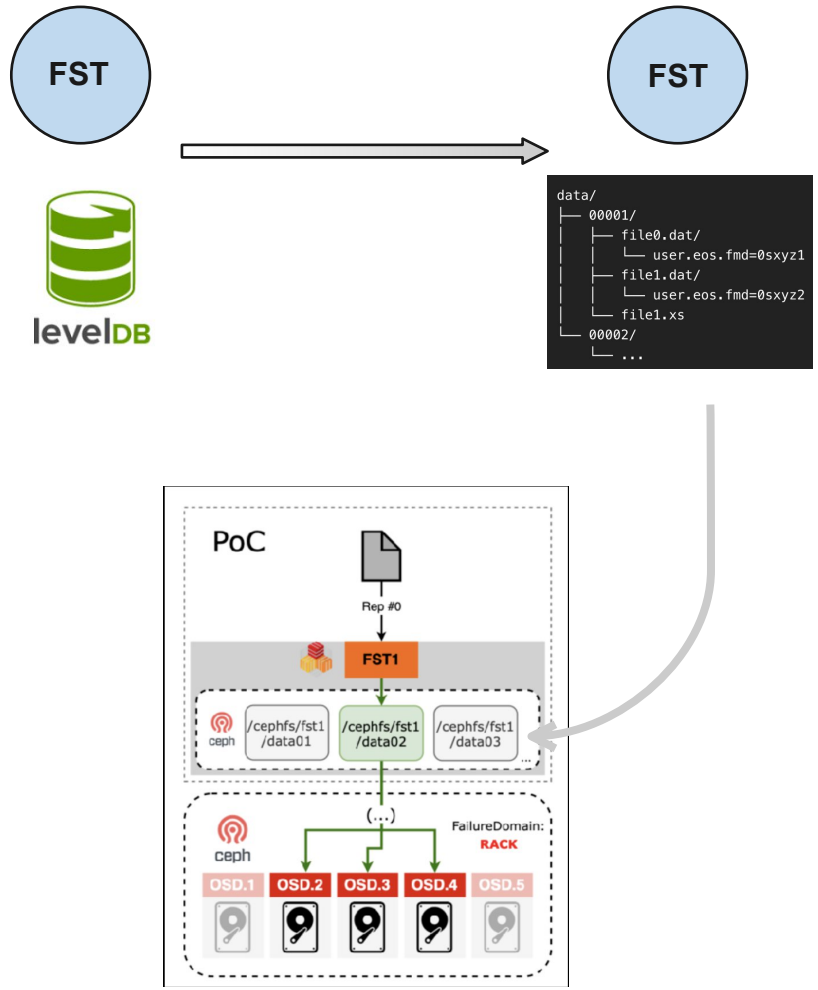
```
[root@eoscms-ns-ip563 (mgm:master mq:master) ~]$ eos io stat -x

┌─> Sum of bytes transferred in last 1m/5m/1h/24h and total sum:

┌──────┬──────────────────────────────────────┬──────┬──────┬──────┬──────┬──────┐
│io    │                           application│  1min│  5min│    1h│   24h│   sum│
└──────┴──────────────────────────────────────┴──────┴──────┴──────┴──────┴──────┘
 out                        config_rem_htautau_2018     0     0     0     0  54.75 M
 out                                        eoscp     0     0     0  9.50 K 573.24 G
 out                             config_diboson_2018     0     0     0     0  73.46 M
 out                              config_dyjets_2018     0     0     0     0   2.74 G
 out                         vhmm_config_diboson_2018     0     0     0     0 204.80 M
 out                                             hadd     0     0     0     0   4.69 T
 out                          nmssm_config_ttbar_2018     0     0     0     0  47.91 M
 out                                             fill     0     0     0     0   5.24 T
 out                         config_ttbar_2016preVFP     0     0     0     0 473.20 M
 out                             Plot_Final_datacalo     0     0     0     0 350.77 K
 out                                        root.exe     0     0     0 123.93 M 715.44 G
 out      tauembedding_tagandprobe_data_2016preVFP     0     0     0     0 391.51 M
 out                    trackPairEfficiencyAnalysis     0     0     0     0  47.69 G
 out                   NanoAODAnalyzer_eltau.exe     0     0     0     0 340.99 M
 out                                config_qcd_2018     0     0     0     0 327.29 M
 out                      nmssm_config_dyjets_2018     0     0     0     0 389.16 M
 out                 nmssm_config_rem_htautau_2018     0     0     0     0   4.34 M
 out                                        python     0     0     0 176.55 K 120.41 G
 out                              config_dyjets_2017     0     0     0     0   1.12 G
 out                             config_diboson_2017     0     0     0     0  40.26 M
 out                              config_ttbar_2018     0     0     0     0   7.09 G
 out                 config_electroweak_boson_2017     0     0     0     0  14.34 M
```

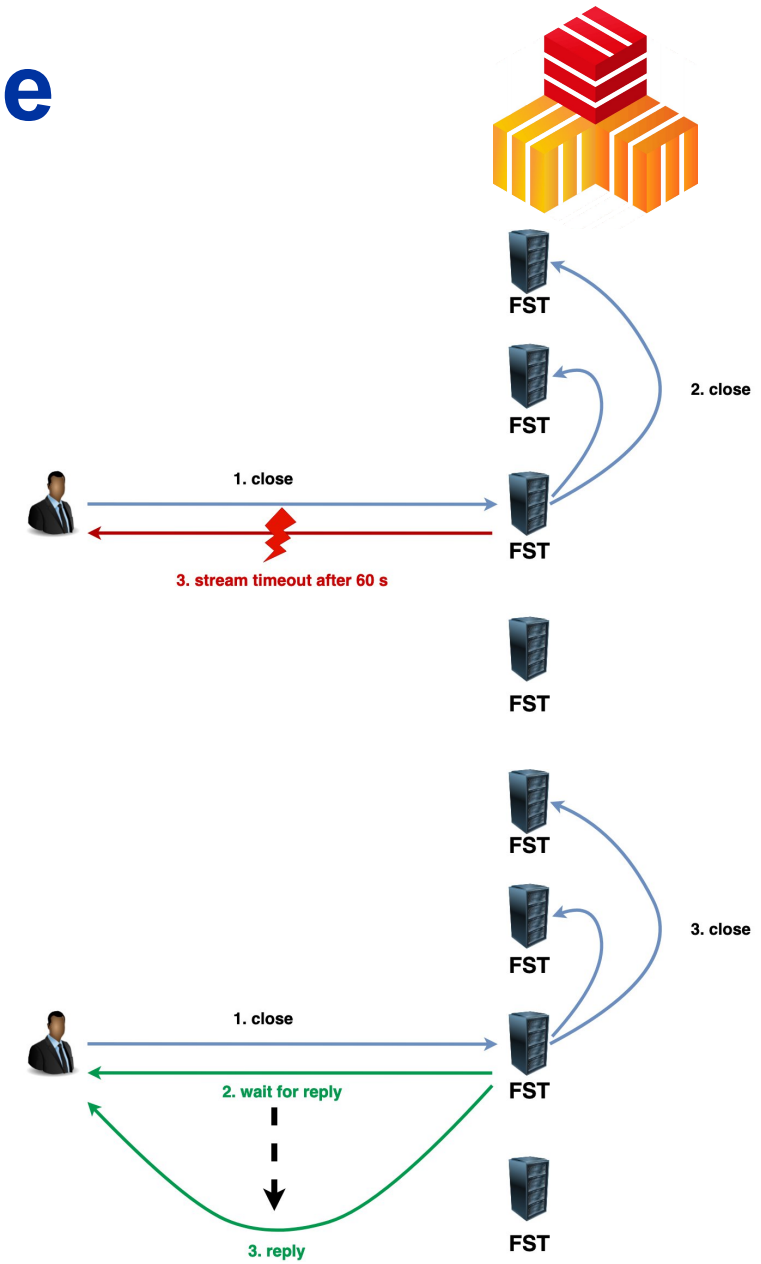
FST developments (1)

- **Move from LevelDB to extended attributes**
  - FSTs are now **stateless**
  - Performance (open times) more **predictable and less jittery**
  - Remove **serialization point on the LevelDB**
- **Re-factoring of the fsck publishing functionality**
  - Publish fsck inconsistencies in **QuarkDB**
  - **Scanner** thread takes care of publishing inconsistencies
  - Remove the need for an **internal publisher thread**
  - **More details**: https://indico.cern.ch/event/1227241/contributions/5349671/
- **Fix and enforce Scanner rate limiting**
  - Relieve some **pressure from the disks**
  - By default 50MB/s can be lowered for disks <= 6TB
- **See talk:** https://indico.cern.ch/event/1227241/contributions/5330894/



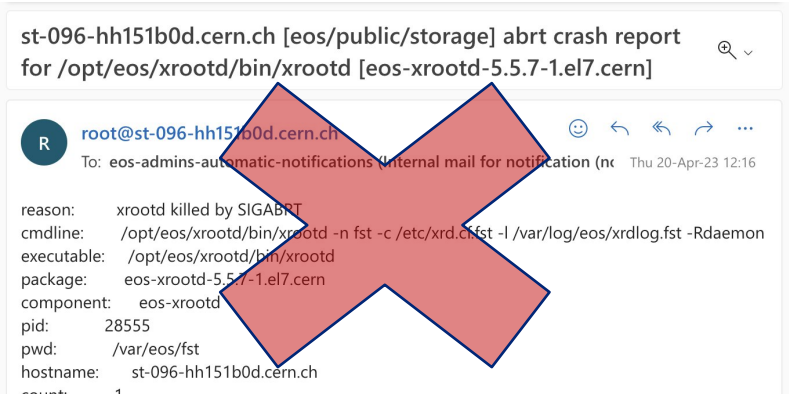
FST developments (2) – async close

- **XRootD async close and the EOS file checksum mechanism**
  - best-effort: for streaming files check is computed "in-flight"
  - for non-streaming cases the file is re-read during the close operation
- **Problem**: for large files (>10GB) `close` take more than the default **XRD_STREAMTIMEOUT**
- **Side-effects**
  - client sees a **timeout error** and a **failed close** operation
  - the server happily re-computes the checksum and closes the file successfully
- **Mitigation**
  - use the **async close functionality (SFS_STARTED / kXR_waitresp)**
  - the client will receive a **notification** from the server then the operation is done
- **Outcome**
  - deployed in production instances and no more complaints from the users
  - many thanks to the XRootD team for fixing a few nasty bug hit along the way
- **Enable** by setting environment variable: `EOS_FST_ASYNC_CLOSE=1`



FST developments (3)

- In shutdown only close FDs we can fsync or that are sockets
  - No more **SIGABRT** crashes during **FST upgrade/restart**
- **Suppress publishing of reports** non-entry RAIN stripes
  - Translates into a **10x-12x reduction** of traffic
- Honor **tpc.ttl** key validity set by clients
  - **Default key validity** extended from 60 to 120 seconds
  - Can be further **adjusted on a per FST basis**
- Add **reporting for slow open operations (>1s)** in the FST logs

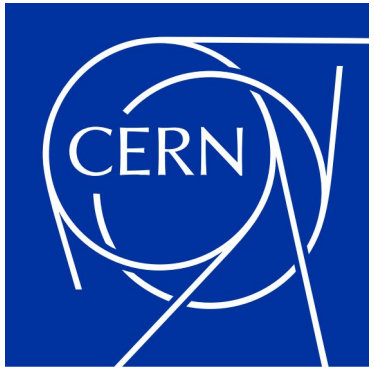



```
230422 04:17:05 time=1682129825.211847 func=open level=ERROR logid=unknown unit=fst@p06636710b70214.cern.ch:1095 tid=00007f49e93f0700
source=XrdFstOfsFile:779 tident=1.37324:306@p06636710b70214 sec=(null) uid=1 gid=1 name=nobody geo="" slow open operation: open-
duration=2364.114ms path='/replicate:ee64e2ef' fxid=ee64e2ef path::print=0.201ms creation::barrier=0.063ms layout::exists=0.007ms
clone::fst=103.314ms layout::open=0.036ms layout::opened=2216.680ms get::localfmd=0.001ms resync::localfmd=0.157ms layout::stat=0.009ms
full::mutex=0.000ms layout::fallocate=0.002ms layout::fallocated=43.517ms fileio::object=0.107ms open::accountingt=0.012ms end=0.008ms
open=2364.114ms
```

Client/FUSE developments

- **Removing** the old **eosd** implementation
  - Starting with release **5.1.1**
- **FUSEX fixes and improvements**
  - Stop file creation earlier for quota or space problems
  - Fix various **recovery scenario**s
  - Effort to avoid **leaving files in inconsistent state -> impact on FSCK**
  - Add **execution time statistics in JSON format**
  - **Rewrite proxy managemen**t using shared_ptr to fix race conditions
- **eosxd3** using **fuse3** implementation
- **eos df command**
  - Abstracts the file layout and provides a **“familiar” view** for used storage
- **eos register command**
  - Allows **arbitrary** modifications for **file meta-data**

Thank you! Questions? Comments?



home.cern