EOS 5 highlights and functionality consolidation

Elvin Sindrilaru

on behalf of the **EOS team**

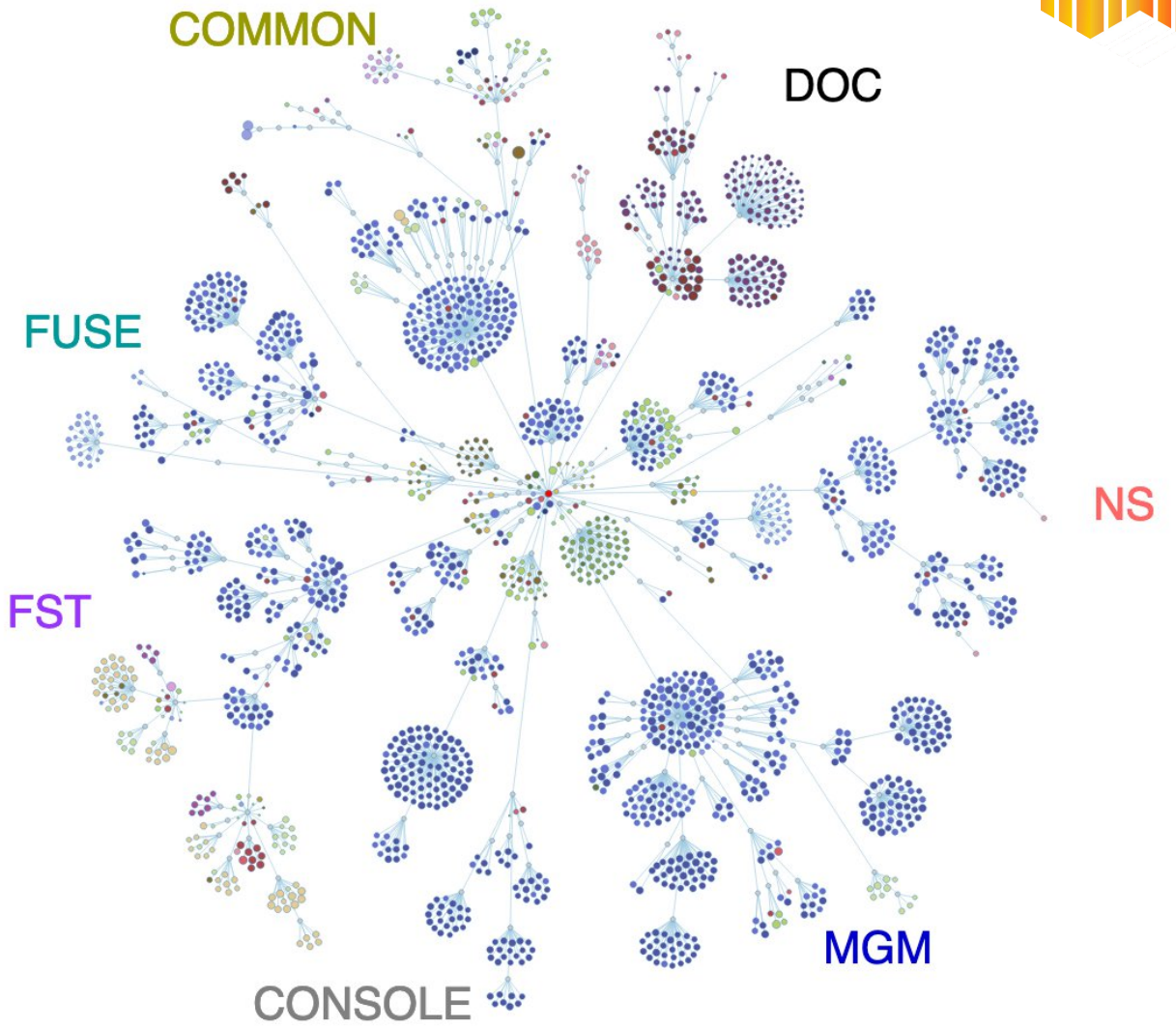
07.03.2022

Outline

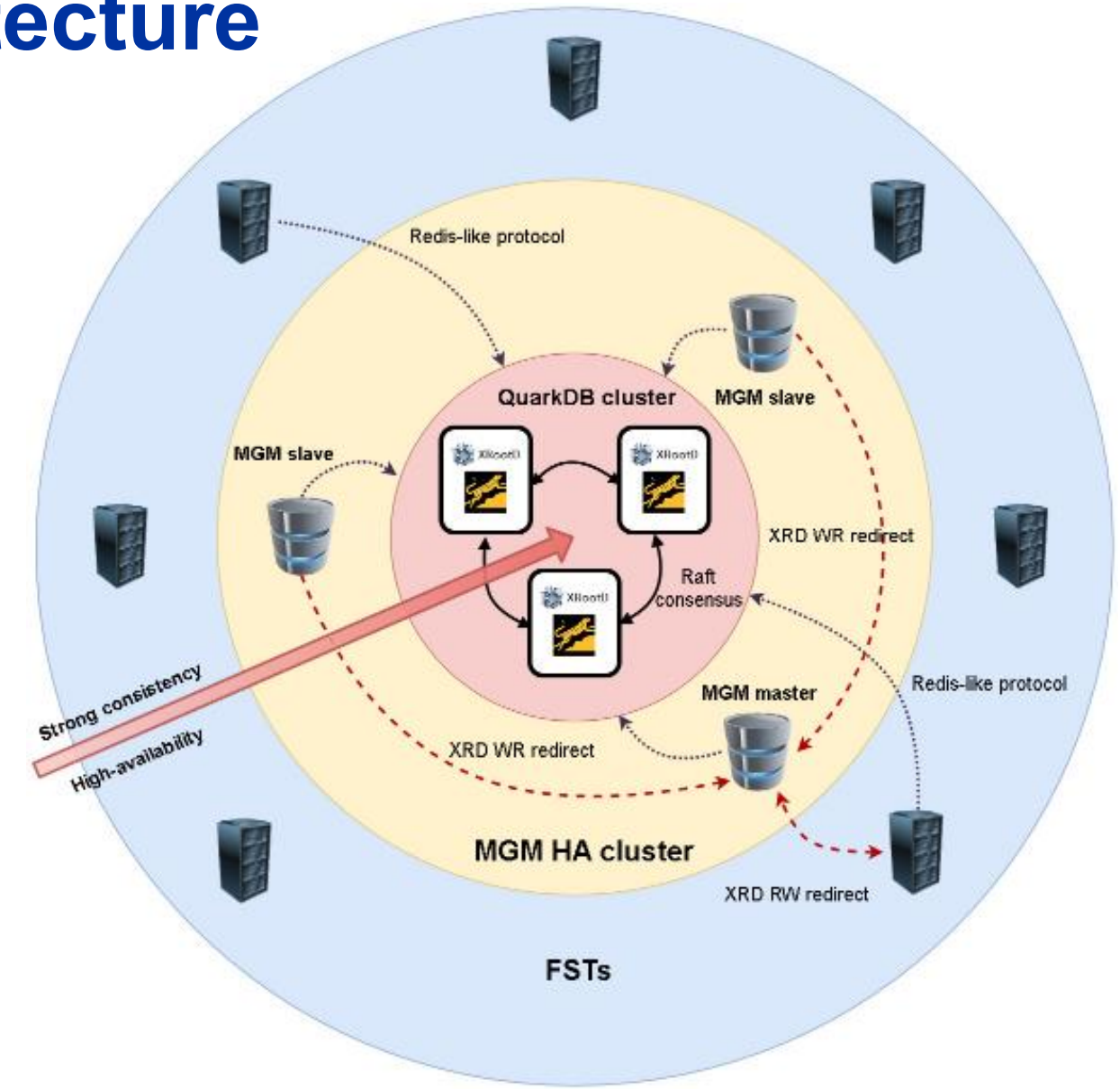
- Removed/consolidated functionality
  - In-memory namespace / File based configuration / Master-slave HA
- MGM improvements
  - IoStat moved to QuarkDB
  - GroupBalancer improvements
  - Rate limiting of client requests
- Support for encryption/obfuscation
- FST improvements
  - Fair scheduling and I/O prioity, direct I/O
  - Asynchronous open/write API
- IAM integration and helper tools
- External contributions
- Plans for the future

EOS 5 in numbers

- Tags: 14
- First release: June 2021
- Latest release: EOS 5.0.13
- Using XRootD: 5.4.2
- Targeted platforms:
  - CentOS 7
  - CentOS 8 Stream
  - (CentOS 9 Stream)
- Best-effort client support:
  - Ubuntu Bionic/Focal
- ASAN builds



EOS architecture

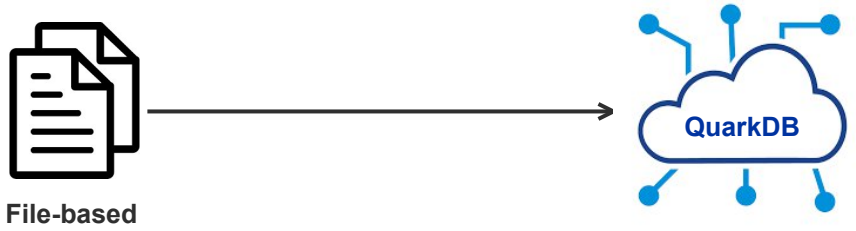


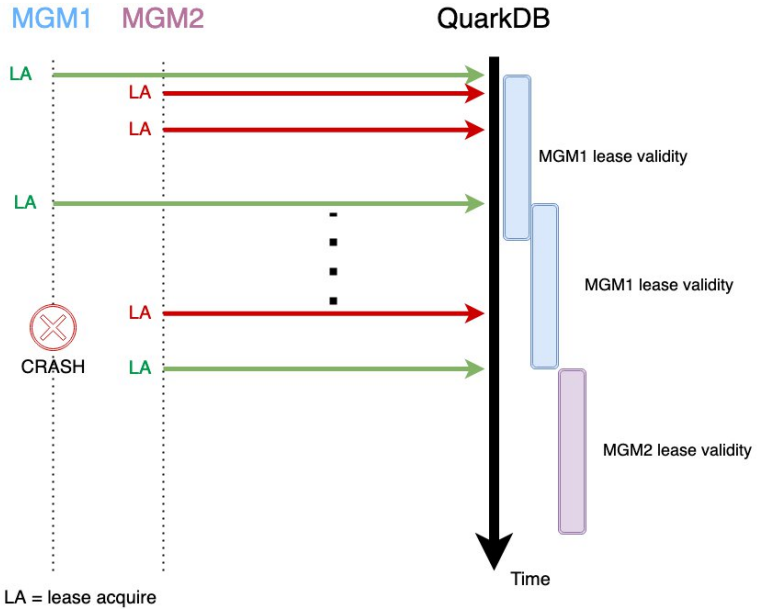
EOS 5 functionality consolidation

- **Namespace** implementation in QuarkDB
  - Dropped support for **In-memory** implementation

- **EOS configuration** by default in QuarkDB
  - Dropped support for **file-based configuration**
  - No more need for running the **eos@sync daemon**
  - **Deprecated** configuration directives:
    - ~~*mgmofs.cfgtype file/quarkdb*~~
    - ~~*mgmofs.autosaveconfig true/false*~~
    - ~~*mgmofs.configdir*~~

- **High-availability** implementation in **QuarkDB using leases**
  - Dropped old HA support





IoStat move to QuarkDB

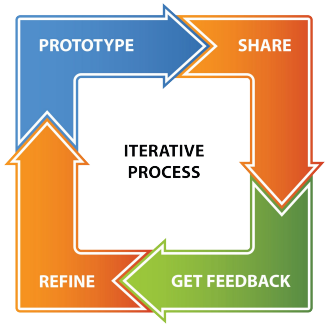
- Complete MGM transition from **stateful** to **stateless**
- IoStat information stored as a **hash-map inside QuarkDB**
  - Automatic migration from file to QuarkDB
- Values are updated **individually and incrementally**
  - Desirable properties in a future sharded MGM setup
  - Using a **Flusher** to push updates to QuarkDB

```
[esindril@esdss000 tmp]$ sudo redis-cli -p 7777 hgetall eos-iostat:eosdev:2022
 1) "idt=g&id=0&tag=bwd_seeks"
 2) "0"
 3) "idt=g&id=0&tag=bytes_bwd_wseek"
 4) "0"
 5) "idt=g&id=0&tag=bytes_deleted"
 6) "21562166829"
 7) "idt=g&id=0&tag=bytes_fwd_seek"
 8) "0"
 9) "idt=g&id=0&tag=bytes_read"
10) "255356046"
11) "idt=g&id=0&tag=bytes_written"
12) "72426072"
13) "idt=g&id=0&tag=bytes_xl_bwd_wseek"
14) "0"
15) "idt=g&id=0&tag=bytes_xl_fwd_seek"
16) "0"
17) "idt=g&id=0&tag=disk_time_read"
18) "181"
19) "idt=g&id=0&tag=disk_time_write"
20) "45"
```

- Important data source for **Grafana plots**
- On-going rewrite of the implementation to improve **accuracy and correctness**
- See related talk by Jaroslav Mon 11:05

Other MGM improvements

- **Rate limiting rules** for various operations
  - **Protect** against one client **grabbing** most of the **resources**
  - Extend implementation to **cover recursive operations**
  - Highly **iterative** process
- **Scalability/locking improvements** related to Fusex
- **GroupBalancer** improvements and extensions
  - See talk by Abhishek Tuesday 9:55
- Move **deletion reports out of MQ**
  - Hitting scalability limits especially for RAIN
- Better handling of **bulk configuration** updates




```
[root@eospublic-ns-ip563 (mgm:master mq:master) ~]$ eos access ls
# ....................................................................
# Banned Hosts ...
# ....................................................................
[ 01 ] googlebot.com
# ....................................................................
# Stall Rules ...
# ....................................................................
[ 01 ]               rate:user:*:Chmod => 500
[ 02 ]               rate:user:*:Chown => 500
[ 03 ] rate:user:*:Eosxd::ext::0-STREAM => 10
[ 04 ]   rate:user:*:Eosxd::ext::CREATE => 250
[ 05 ] rate:user:*:Eosxd::ext::LS-Entry => 10000
[ 06 ]      rate:user:*:Eosxd::ext::SET => 50
[ 07 ]   rate:user:*:Eosxd::ext::UPDATE => 250
[ 08 ] rate:user:*:Eosxd::int::FillContainerMD => 10000
[ 09 ]      rate:user:*:Eosxd::prot::LS => 1000
[ 10 ]            rate:user:*:GetFusex => 200
[ 11 ]           rate:user:*:OpSetFile => 300
[ 12 ]               rate:user:*:Open => 500
[ 13 ]       rate:user:*:OpenDir-Entry => 20000
```

Support for encryption/obfuscation

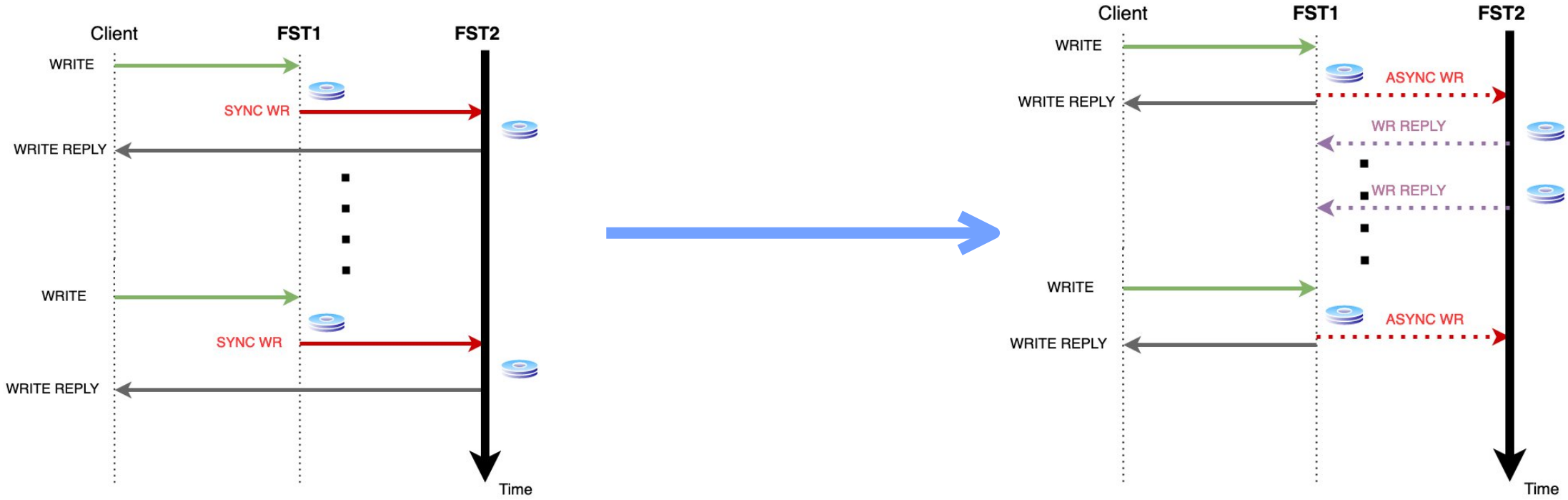
- Encryption at the transport layer thanks to **XRootD 5**
  - Server decides to what extent encryption is enforced
    - default: **off**
    - enforce if client supports: **capable**
    - meet client's requests: **none**
  - Pre-requisite for **token authorization**
  - Possibility to **encrypt only meta-data operations**

- **Unified and shared TLS** support for **xroot** and **http** protocols

- **Encryption**/obfuscation of the **data stored on disk**
  - See talk by Andreas Tue 11:10

FST I/O performance optimizations (1)

- Implement **asynchronous open operations**
  - **Considerable impact for RAIN layouts** with many stripes
- Implement **asynchronous write operations**
  - General speed-up of transfers due to **pipelining**
  - All previous **consistency guarantees still hold**
  - Considerable impact in **high latency FST setups**



FST I/O performance optimizations (2)

- **Early detection of errors** to avoid wasting bandwidth and resources

- **Buffer pool implementation** used for all I/O operations
  - Asynchronous mode comes with more memory requirements
  - All buffers are **memory aligned** to ensure possibility of doing **direct I/O** (O_DIRECT)

- **Scheduling limits for file systems** to avoid hot-spots
  - *max.ropen - maximum number for read streams per file system*
  - *max.wopen - maximum number of write streams per file system*

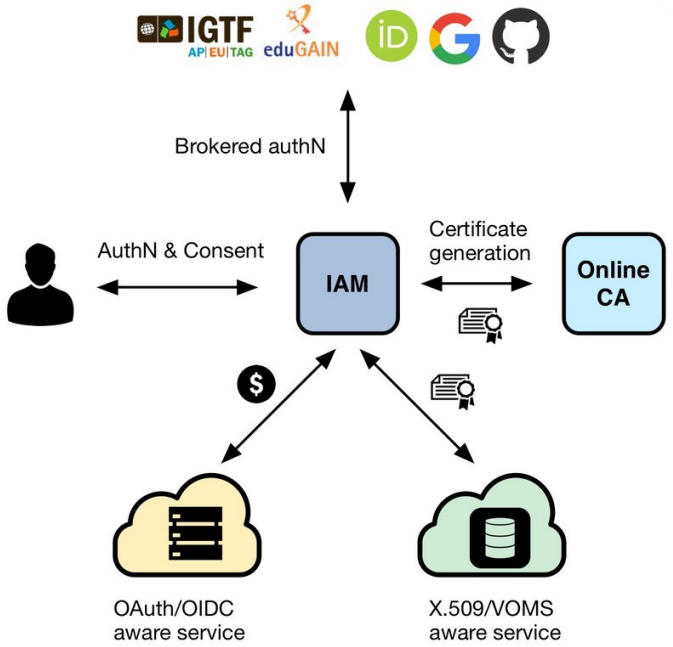
- Allow tagging each **transfer with a certain I/O priority**
  - Defined at the **space level:** ***space config <name> space.policy.iopriority=<val>***
  - **Per transfer** using opaque info: ***xrdcp root .... ?eos.iopriority=<val>***
  - Follow the **Linux ionice convention**: **0 (high priority) - 7 (low priority)**

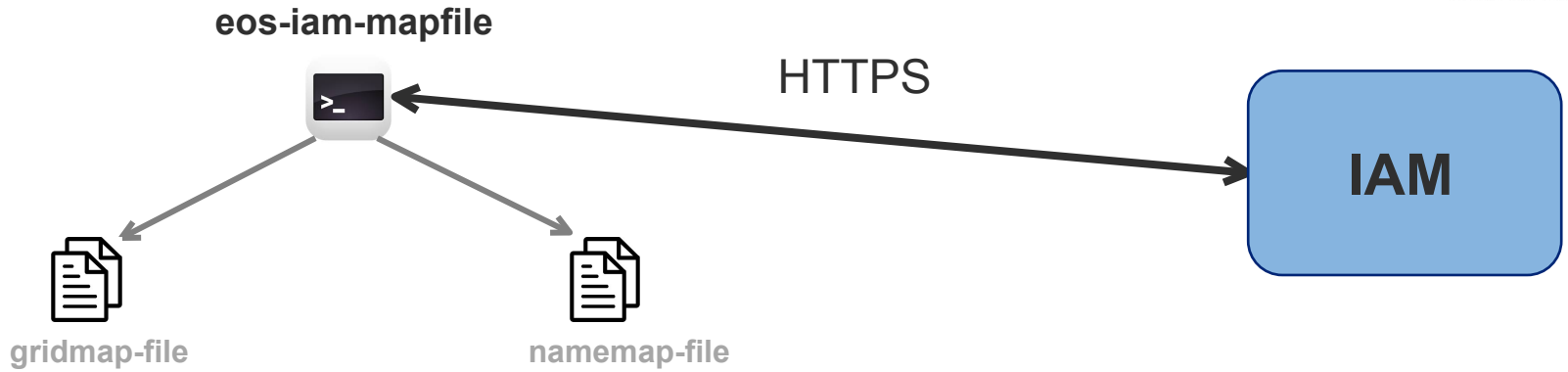
- Add support for **XRootD pgRead/pgWrite** API

IAM integration and helper tools

- Experiments are planning to **decommission** the **existing VOMS infrastructure**
- Dedicated tool to build **gridmap-file** and **namemap-file**
- ***eos-iam-mapfile*** tool
  - Connects to **IAM (Identity and Access Management)** providers
  - Acts as a replacement for ***edg-mkgridmap*** tool
  - Construct **gridmap-file** used by existing GSI plug-ins

    <user_DN> local_username
  - Construct **namemap-file** used to map tokens to local identities





External contributions

- **EOS Windows Native Client developments**
  - Feature-rich EOS WNC
  - Move most of the EOS console commands to ProtoBuf
  - See dedicated talk Tue 11:45

- **University of Vienna - E. Birngruber**
  - Support health status reporting for Linux multipath

- **Reykjavik University - J.T. Foley**
  - Documentation improvements

- **Australia's Academic and Research Network**

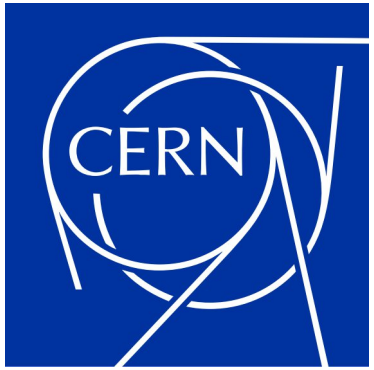
Plans for the future

- More **functionality consolidation**
  - Drop support for **eosd**
  - Drop support for **libmicrohttpd** and rely on **XrdHttp**
  - Drop support for the internal **Transfer Engine** - anyone using this?

- **Balancing** rewrite to use native **XRootD Third-Party-Copy operations**
- **Drop the MQ daemon**

- **Stateless FSTs** by dropping the local LevelDB
- Consolidation of the **internal I/O API** - once other dependencies are dropped

- Looking forward to more **input from the community!**



home.cern