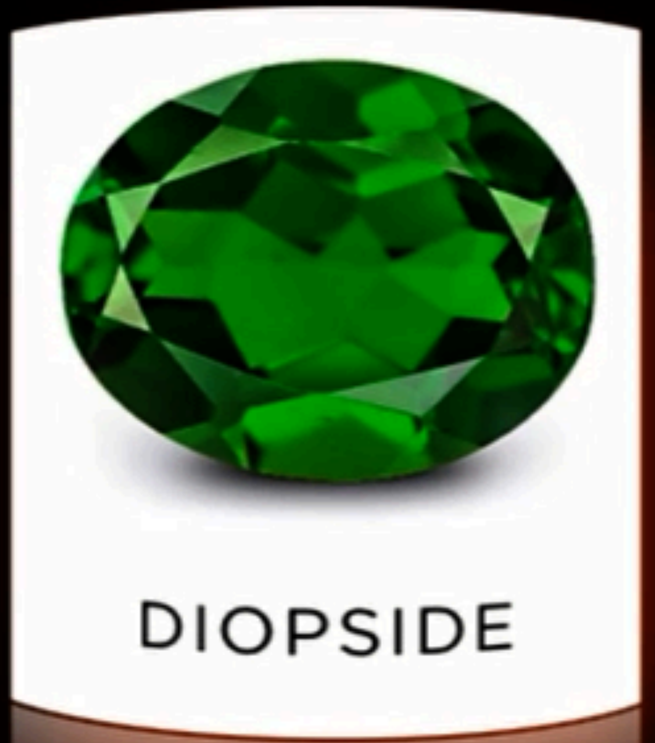
EOS 5 Roadmap - Run 3

EOS Workshop 2022

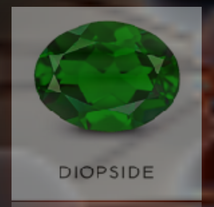
Andreas-Joachim Peters
CERN IT-ST for the EOS project

EOS Open Storage



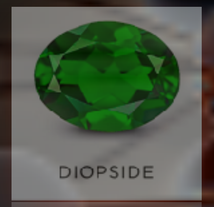
Production Status of **EOS** 5 .. and where do we go …

EOS 5 & XRootD5 Production Status



- **EOS5** tested **XRootD5** to the teeth …
  - very complex use-case of **XrdCl** inside **eosxd** and **MGM/FST/MQ**
  - the differences in the code base of **EOS4** vs **EOS5** internally are moderate
    - more intrusive changes inside the XRootD5 framework - new APIs …
- as of today we have only one instance with EOS5 **eosams02**
- we don’t have yet **EOS5** clients rolled out on **lxplus/lxbatch** in production - **why**?
  - we had several iterations of EOS5 **in QA** but after several updates rolled-back **QA**
    - regressions/bugs found in **XRootD5**
    - regressions found in eosxd recovery when moving from codebase **EOS 4.8.51 -> 5.0.9**
  - since we use XrdCl heavily inside the server **we were cautious** to put EOS5 server into production before we are convinced about the client functionality

EOS 5 & XRootD5 Production Status



- **EOS5 5.0.14** is on the way with XRootD **5.4.2** and will then be rolled-out **ASAP** client-side on lxplus/lxbatch using the usual QA chain and then to production
- for a better understanding **5.0.13** is usable - however for *extreme* client usage it has known problems, which may be triggered in high concurrency/load situations
- we will in the next weeks stop supporting the **4.8.x** release branch
  - we will still provide critical bug fixes on request

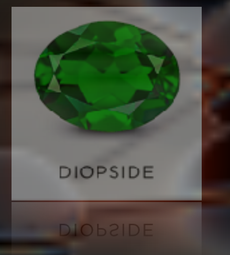
EOS 5 Deprecations

- ongoing: **remove dead weight** from the past and simplify EOS
- we will **drop** **libmicrohttpd** (http) and have https using **XrdHttp** as default
- we **dropped** the **in-memory** namespace implementation and provide the **QuarkDB** namespace implementation only
- we **dropped** *BERYL* **HA setup** with master/slave/sync process
- we will **drop** **support** of eosd and will **remove** it completely in **5.1**
  - if you still use it - time to move to eosxd
- we want to **drop** the **MQ** service for messaging and use QuarkDB pub-sub as drop-in replacement in **5.1++**
- we **dropped** MGM **configuration files** and provide configuration **in QuarkDB** only
- we **dropped** local IOSTAT files and store **in QuarkDB**

EOS 5 Namespace Locking

- boost the namespace performance using **shared_ptr** and **object local locks/atomics** from version **5.1** on
  - we have already many long lasting lock fixes in 5.0
  - local lock strategy already successfully inside **eosxd**
  - the performance of the fair *SharedMutex* is not fantastic - in particular under contention - so avoid it where possible
    - actually: FUSE is access by inode - this opens the possibility to not use the hierarchical global lock in the FuseServer implementation and avoid most of the contention - currently versioning is done using logical paths and would require this lock
- reduce latency tail and thread pile-up effects
  - better thread scaling - MGM service scale-up with high CPU core trend
- single MGM optimisations are more promising than a split of the MGM service
  - splitting includes an additional operational and hardware cost and might give you only a factor 2,4,8…
  - logical subtree-splitting does not protect from the usual overload scenario where a single user works in a subtree or single directory with several thousand clients

EOS 5 Stateless FST



- FST stores a copy of some relevant file meta-data in LevelDB
- had several attempts already to remove LevelDB from FST and store information as extended attributes - it has to happen **now**
- … because LevelDB is an unpredictable source of latencies on FSTs
  [and outdated and …]

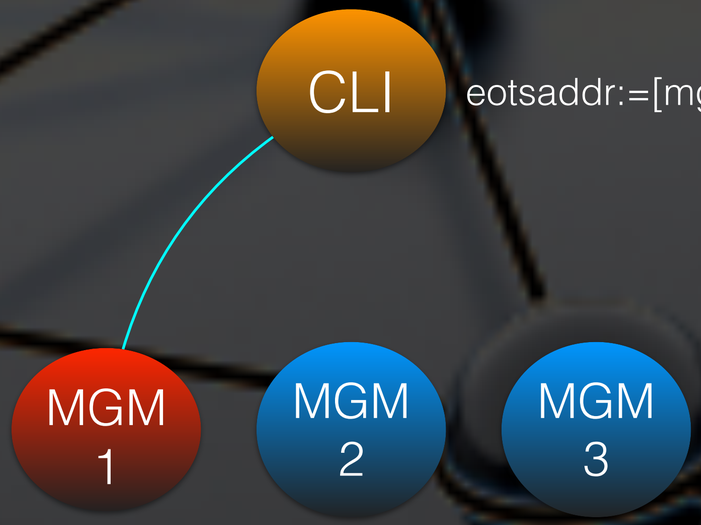
```
220309 03:44:03 time=1646793843.636488 func=open                    level=INFO  logid=c824bc58-9f52-11ec-82c8-a4bf0162970b
unit=fst@p05151113071960.cern.ch:1095 tid=00007f07c9dfc700 source=XrdFstOfsFile:698             tident=root.18923:155@eosmon01 sec=(null) uid=18118
gid=2688 name=nobody geo="" open-duration=58.599ms path='……..' fxid=00495f69 path::print=0.310ms creation::barrier=0.087ms layout::exists=0.008ms
get::localfmd=40.012ms resync::localfmd=0.267ms clone::fst=0.001ms layout::open=0.019ms layout::opened=28.731ms layout::stat=0.009ms
full::mutex=0.000ms layout::fallocate=0.002ms layout::fallocated=29.080ms fileio::object=0.050ms open::accountingt=0.017ms end=0.006ms open=98.599ms
```

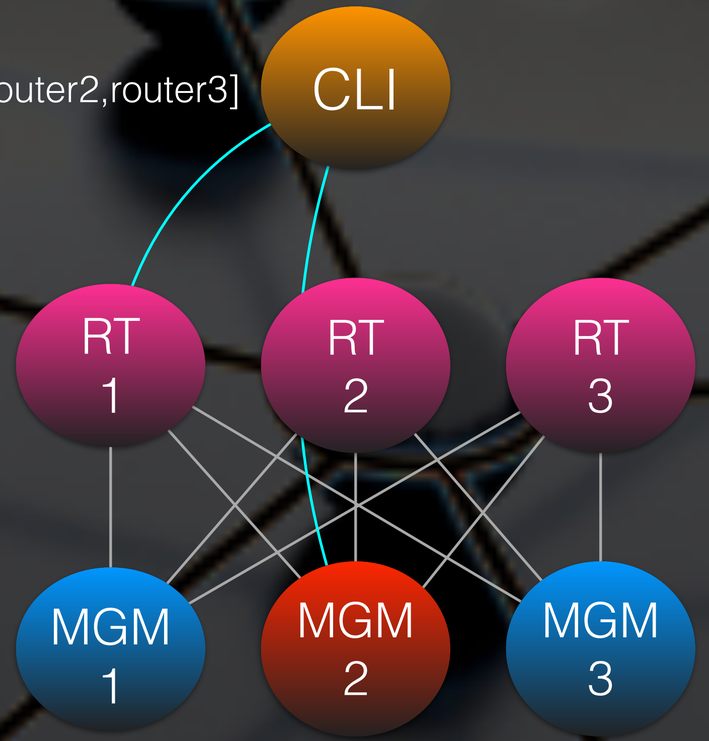
- a completely stateless FST eases complete storage virtualisation and non-POSIX remote storage behind FSTs
- FSCK scans also suffer/introduce locking problems on LevelDB

EOS 5 High Availability

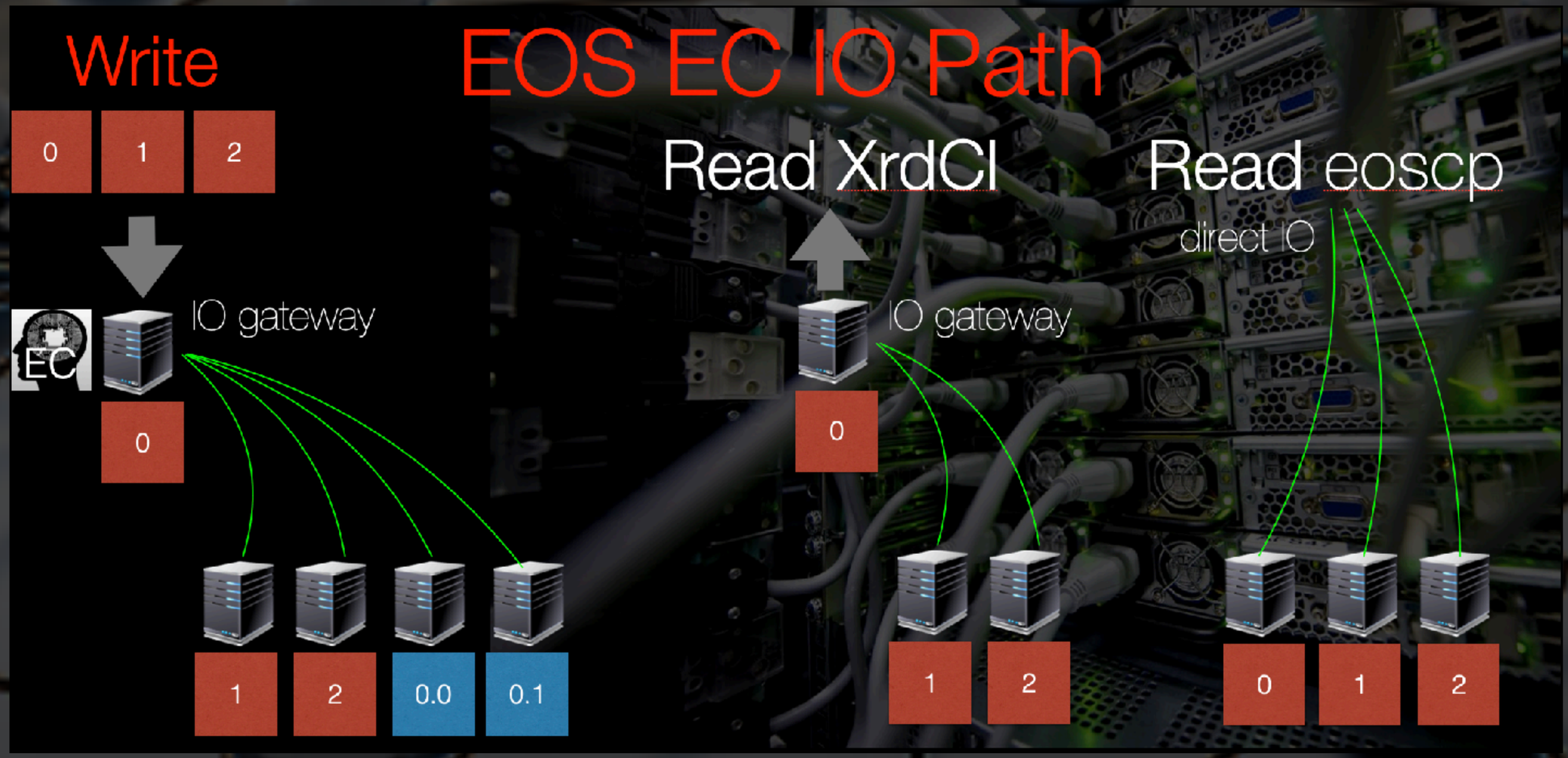
- redirect collapse feature of XRootD5 allows efficient service failover without the need of virtual IPs and/or dynamic DNS entries

  - clients are redirected or connecting to any MGM or ROUTER and they stick to it until this MGM sends them away using a redirect or the connection breaks



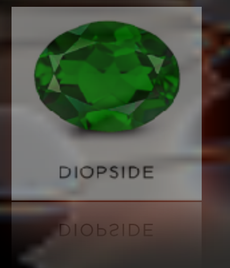


EOS 5 Erasure Coding [1]



EC implementation in **EOS4** and **EOS5**

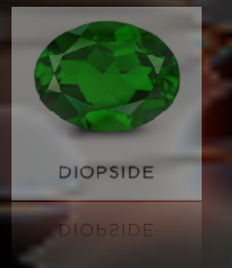
EOS 5 Erasure Coding [2]



- **XrdEc** integration into **EOS5** has been postponed
  - integration of XrdEc into EOS5 **not a low-hanging fruit** including life-cycle management
    - however POC is done
  - **priority** to deliver O2 with well-proven EC implementation into production for Run-3 - done - works well, so why **XrdEc**?

- **client-driven EC** with **XrdEc**
  - read-traffic amplification **1.0** instead of ±2.0 - read latency factor 1.0 instead of 2.0
  - write-traffic amplification **1.x** instead of ±2.0 - but client-server traffic increases!!!
    `x = (m+k)/m    e.g. 1.25 for 8 data + 2 parity disks`
  - client plug-in configuration and roll-out is problematic - very likely that we will use the plug-in mainly inside the FST and only few high-performance clients could benefit from it on client-side
  - best would be to enable from server-side
- still planned (CERN author of XrdEc) : work on new **EOS layout** for **XrdEC**
  - configurable **`RS(m,k,l)`**
    - `m:=data blocks k:=parity blocks l:=parallel diskset`
      - l is an additional parameter describing how many disk get involved for parallel IO - it allows to scale-up per file throughput + IOPS additionally to (m+k)
- we would push for the **EC Update** feature in XrdCl - using copy-on-write mechanism of XFS - because it is one more convincing argument to add **XrdEc** to EOS

EOS 5 Filesystem [eosxd]



- **EOS** FUSE implementation is **THE** crucial interface for users
  - it is **used a lot** at CERN *[40% of read traffic]* because of convenience … although it is often not the most performant way to access data and more sensitive to backend problems / latencies
  - we have the **most versatile authentication system** for a filesystem
  - we have the **most versatile permission and quota system** for a filesystem
    - **it makes a lot of sense to invest further work into the implementation**
- Work to do:
  - **Refactor** XrdClProxy class using declarative XRootD5 API and writeV
    - **eosxd** implements an extended XRootD client class XrdClProxy
      - this adds dynamic+static read-ahead features and simpler async interface
      - with refactoring a lot of complexity can be removed from eosxd
        - will make the implementation more robust against some rare race conditions
  - **eBPF** - extFUSE model [try collaboration with CVMFS …]
    - eBPF allows an application to share e.g. maps with meta-data between a user-space process and kernel module
    - allows to avoid many round trips between kernel FUSE and FUSE process

EOS 5 SquashFS Feature

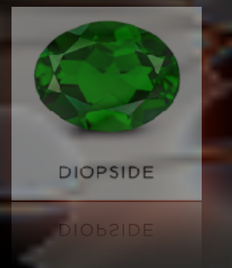
- **EOS**/FUSE is not the filesystem of choice for software compilation and distribution
  - compilation is an extremely latency accumulating use case
  - we have a neat solution which unfortunately is not available in production
    - **SquashFS** functionality
      - allows mounting and using read-only software distribution images
      - contains a complete **release management interface** inside the EOS shell

Using SquashFS images for software distribution

EOS provides support for SquashFS image files, which can be automatically mounted when the image path is traversed. This functioality requires an appropriate automount configuration.

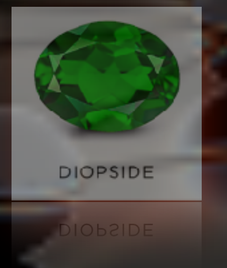
https://eos-docs.web.cern.ch/using/squashfs.html

EOS 5 SquashFS Feature



- **We** will push **SquashFS** functionality into QA at CERN when EOS5 clients are in production
  - it is *just* an additional package with automount reconfiguration
  - we might configure some dedicated space for software packages offering high-throughput to a larger number of clients to access image files

EOS 5 Distributions



- **EOS5** is distributed within EGI
- currently we provide **server RPMs** for CentOS7/8S
  - we plan to provide more platforms (e.g. CentOS9, Ubuntu & ARM) during 2022

EOS 5 Timeline

EXPECT DELAYS

Stable **Production Version**
for **Clients**
for **Server**
5.0.14+

Workshop 2021	Q2/21	Q2/22	Q3/22	....
5.0 beta	**5.0**		**5.1**	**5.2**
XRootd5 Deprecations New Configuration	More Deprecations New Configuration Started MQ Transition EGI CS8 Stream		*New* Locking Share ACLs Encryption **eosxd** using new API drop eosd *New* HA Stateless FST	XrdEc Layout (Updates?) eBPF **eosxd** prototype No MQ

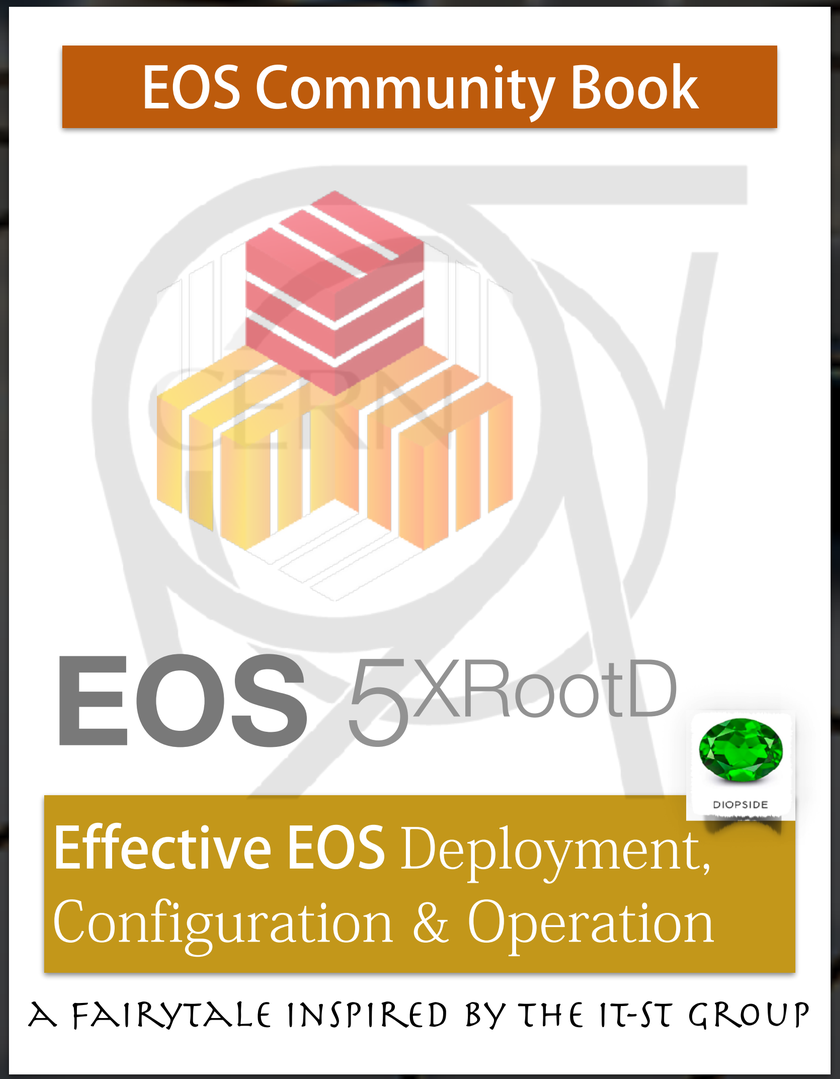
https://gitlab.cern.ch/dss/eos/-/tree/eos5

EOS 5 Community

- **interesting projects** to participate as members of the **EOS** community
  - Contribute migration tools from storage *abc* to EOS
  - Join an **EOS** documentation effort
  - Share your story in the **EOS** community
  - Contributions to **Storage Virtualisation** (Kubernetes++, shared FS)
  - Packaging/Configuration/Virtualisation of **multi-user MinIO S3** front-end service
  - **Optimisation** of EC configurations for physics use cases
    - TTree and RNTuple - support record & replay / collection & running of samples
  - **Kernel FUSE** improvements for **eosxd -** eBPF
  - and more …

I think: **It Is Time for a Trilogy**

EOS 5 Manual I



CTA Manual II

CTA Community Book

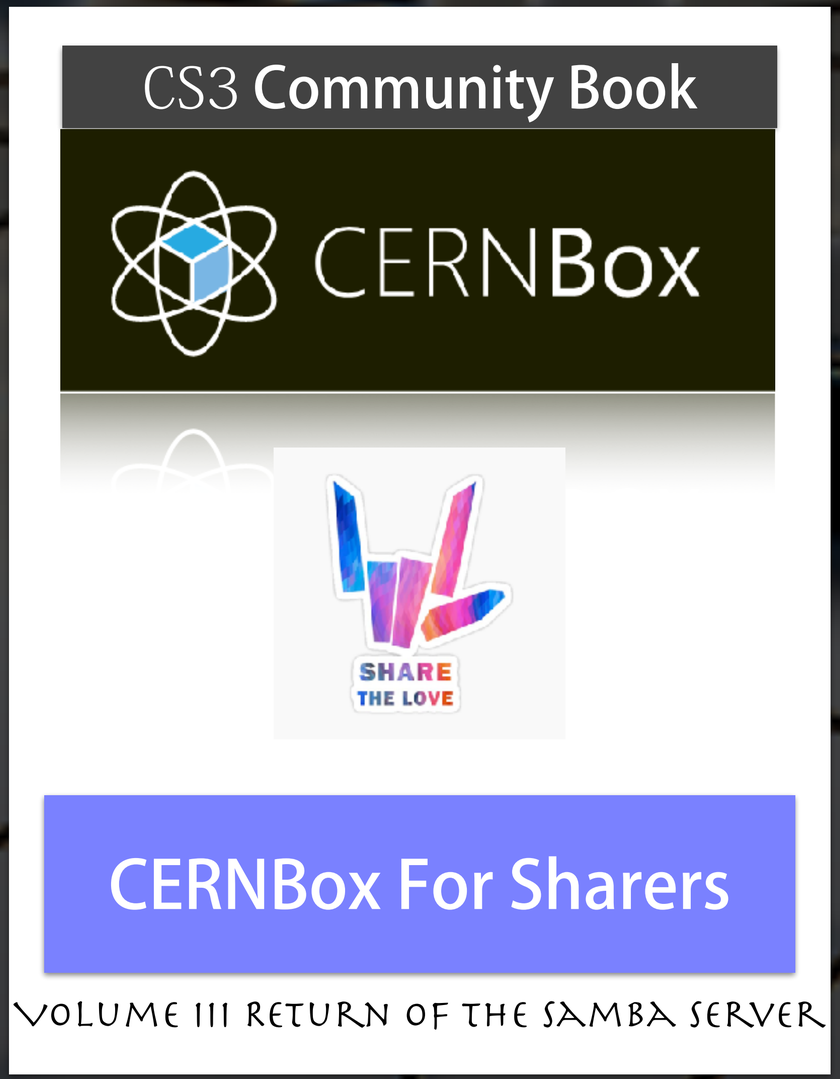
CERN
Tape Archive

CTA

CTA For Dummies

VOLUME II THE TAPE SPOOLING CONTINUES

CERNBox Manual III



EOS Page

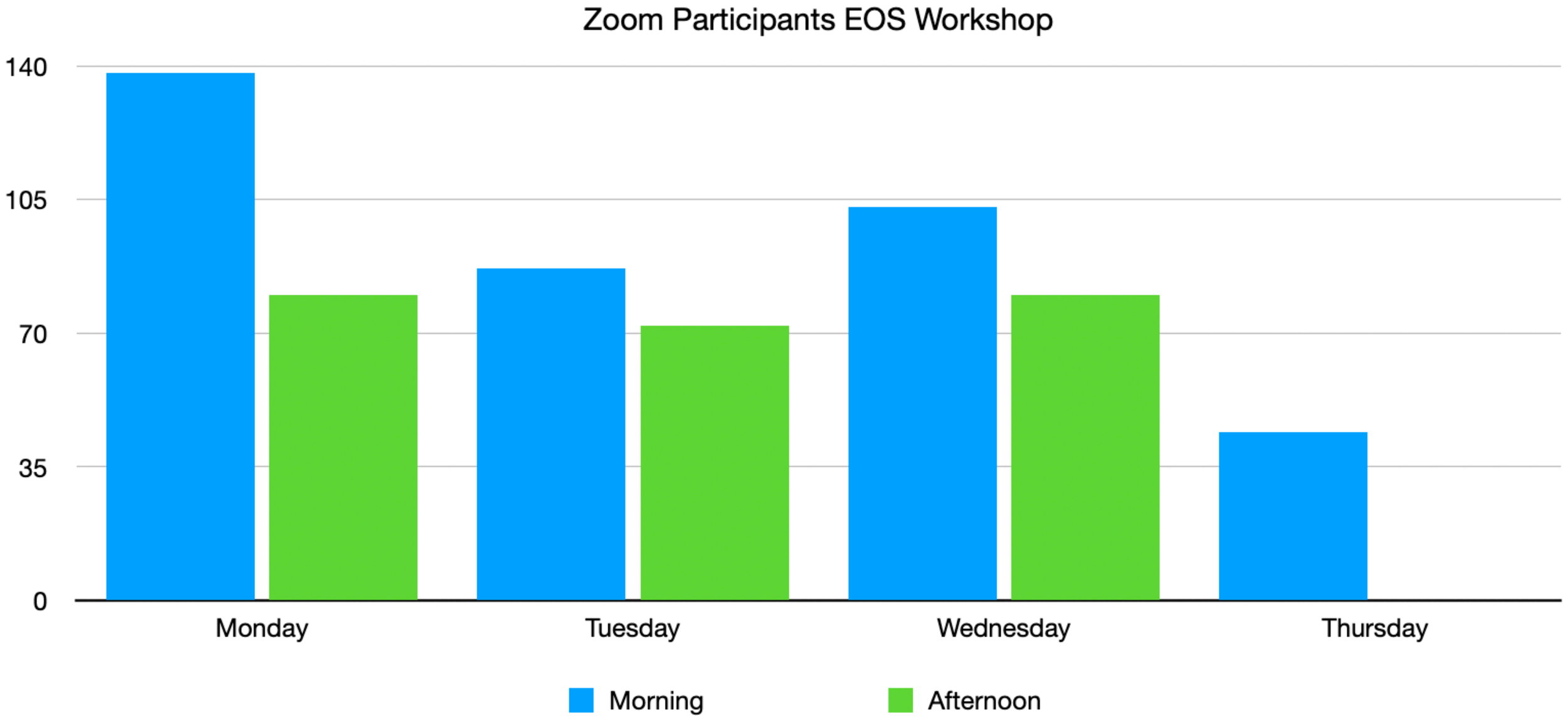
… will be updated with state of the art documentation for EOS 5.0

my EOS Vision

- it is **not so important how** you deliver **functionality & stability & continuity**, it is the **functionality, stability & continuity** itself which **is most important**
  - we might at some point exchange, stack and/or remove building blocks of **EOS**
    - **EOS** repesents a modular storage system with flexible configuration and: the *best* configuration to chose can be different for the person/service looking at it
- **EOS** stands for a successful, long-running **Open Source** project at CERN
  - CERN relies on it heavily and supports it for **Run-3++**
  - it is a safe bet to follow CERN - if we fail others won't have much to store …
- You know: "**Storage is quite stateful**" …
- … especially *hundreds of PBs* and billions of files are …

EOS Summary & Outlook

- A big **THANK YOU** to all of you for your participation, interest and fantastic contributions - and for very constructive feedback added in your contributions
- We hope that you enjoyed the contributions covering the four workshop areas
  EOS, XRootD, CERNBox & CTA
- **All presentation slides are available** in INDICO - the recordings of all presentations should be available with 1-2 days delay - many are already added
- The **CERN team is really glad about the community work** and enjoys to help wherever they can - don't hesitate
  to contact us for any question or problem you are facing or to share your expertise with us
- We really hope that we can do **the 7th workshop in 2023** as a real event to add back a very important part: the social networking of the community in the same room
- **You will not see a summary of the workshop now** - very likely we will provide a summary in a short while and present it at various meetings/workshops/conferences


Zoom Participants EOS Workshop
140
105
70
35
0
Monday
Tuesday
Wednesday
Thursday
Morning
Afternoon