



CERN

European Organization for Nuclear Research  
Organisation Européenne pour la Recherche Nucléaire



# EOS- Disk Storage at CERN

Andreas-Joachim Peters  
IT-DSS

Acknowledgements for participation, help, contributions & discussions to IT-DSS & IT-ES Group, XROOT project & ATLAS & CMS team et al.

ACAT 2011 - London

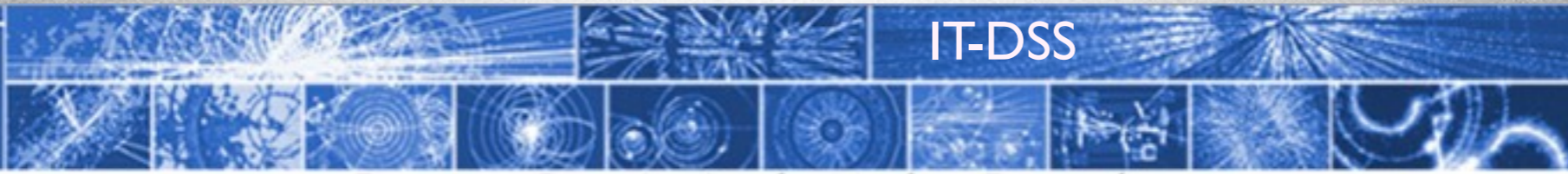
[andreas.joachim.peters@cern.ch](mailto:andreas.joachim.peters@cern.ch)

CERN IT Department  
CH-1211 Genève 23  
Switzerland  
[www.cern.ch/it](http://www.cern.ch/it)

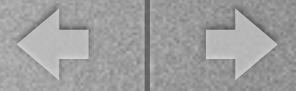


# Outline

- Introduction - “What is EOS”
- EOS Version 0.1.0
- EOS Production Instances
- EOS Operations at CERN
- Roadmap/Outlook



# Introduction to EOS



# EOS Disk Pool Project

- Started after project mandate in April 2010 in IT-DSS with storage architecture discussions with small team
- Since **May 2010** 1<sup>st</sup> development phase
- Since **August 2010**
  - Evaluation in **LST 2010** with ATLAS (**L**arge **S**cale **T**est - 1.5 PB pool)
- **Jan-April 2011** - Upgrade of core communication (shared hashes/queues)
- Since **May 2011** - Production instances for CMS & ATLAS

**GOAL** => Migrate disk-only activity from CASTOR to EOS for optimized resource usage in CASTOR & EOS



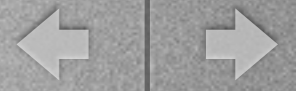
# What is it ...

- **Easy to use** standalone **disk-only storage** for user and group data with in-memory namespace  
(only few **ms** read/write open latency)
- **filling a gap** between AFS (kb Files) and MSS (large file streaming e.g. CASTOR)
- based on **XROOT server plugin** architecture
- **merging ideas** from Hadoop, XROOT, Lustre et al.
- not solving all possible use cases  
e.g. no MSS - complementary to CASTOR
- **fitting to CERN** hardware  
(low cost hardware - no high-end storage)



# Some Requirements ...

- **POSIX like rw** file access (random + sequential + update)
- **Hierarchical Namespace**
  - $10^8$  files [achieved with 128GB memory]
  - $10^{6-7}$  container(directories)
- **Strong Authentication, Quota, Checksums**
- **High Availability**/redundancy of services & data
- **Dynamic** pool hardware  
**scaling & replacement** without downtimes
- ...



# Why EOS and not ...

- ***CASTOR***

- complex system designed for T0/CDR use cases
  - in conflict with other use cases e.g. what is good for analysis is not for CDR

- ***LUSTRE***

- not recommended (yet) after evaluation in 2010 by CERN team

- ***dCache***

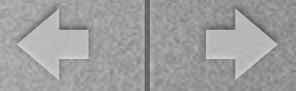
- similar focus as CASTOR on MSS functionality

- ***Hadoop/XROOT/DPM***

- *requirement mismatches*

Move away  
from HSM model!

Introduction



# Access Protocol

- EOS uses **XROOT** as primary file access protocol
- **XROOT** protocol leaves more flexibility for enhancements than NFS4 protocol - but not a design limitation (could be changed)
- protocol choice is not the key to performance as long as it implements the required operations, **but**
  - **SERVER**: data delivery is limited by disk IO + network bandwidth using XROOT protocol - true also for *http*, but not for *HADOOP*
  - **CLIENT**: Caching matters most
    - currently XROOT client is not ideal concerning the caching (rewrite started ...)
    - on the contrary XROOT protocol via a FUSE mount shows identical performance as an NFS4 or Lustre mount for most use cases





# Architecture

## Management Server

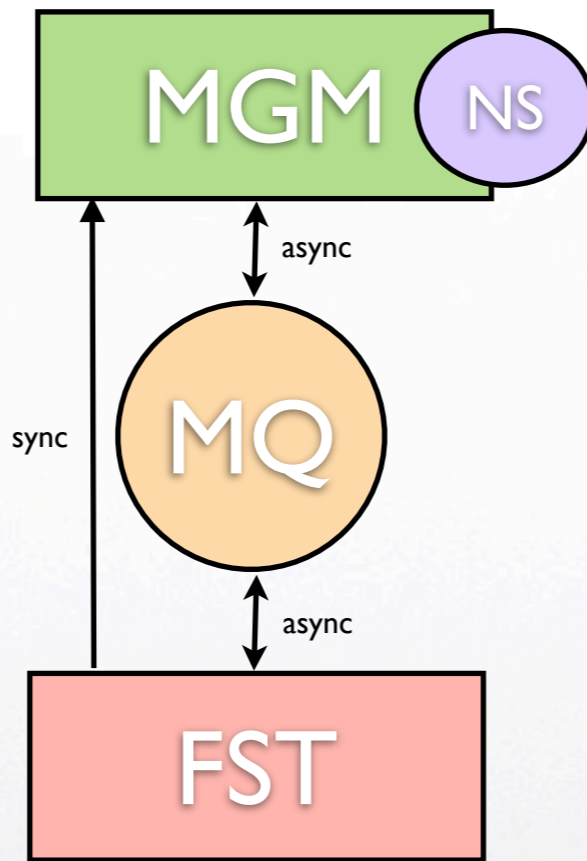
- Pluggable Namespace, Quota
- Strong Authentication
- Capability Engine
- File Placement
- File Location

## Message Queue

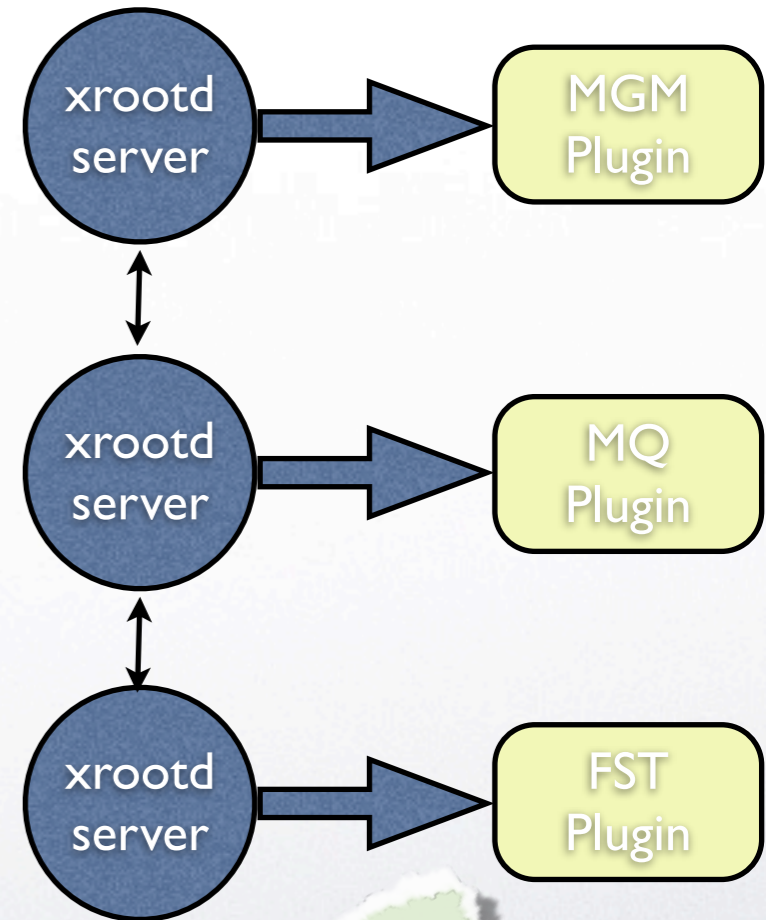
- Service State Messages
- File Transaction Reports
- Shared Objects (queue+hash)

## File Storage

- File & File Meta Data Store
- Capability Authorization
- Checksumming & Verification (adler,crc32[c],md5,sha1)
- Disk Error Detection (Scrubbing)



Implemented as plugins in **xrootd**



No DB Backend required!

Introduction



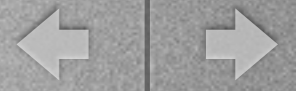
# Additional Services

- **FTS/GRiD Access Point**

- **BestMan** SRM running on EOS-Fuse mount point
  - only 1 Hz file creation rate :-)
- **gridFTP** with EOS-DSI plugin using xrootd Posix
  - ATLAS successfully used FTS without SRM on target end
    - file creation bottleneck removed on target side

- **EOS Sync**

- changelog & configuration file replication



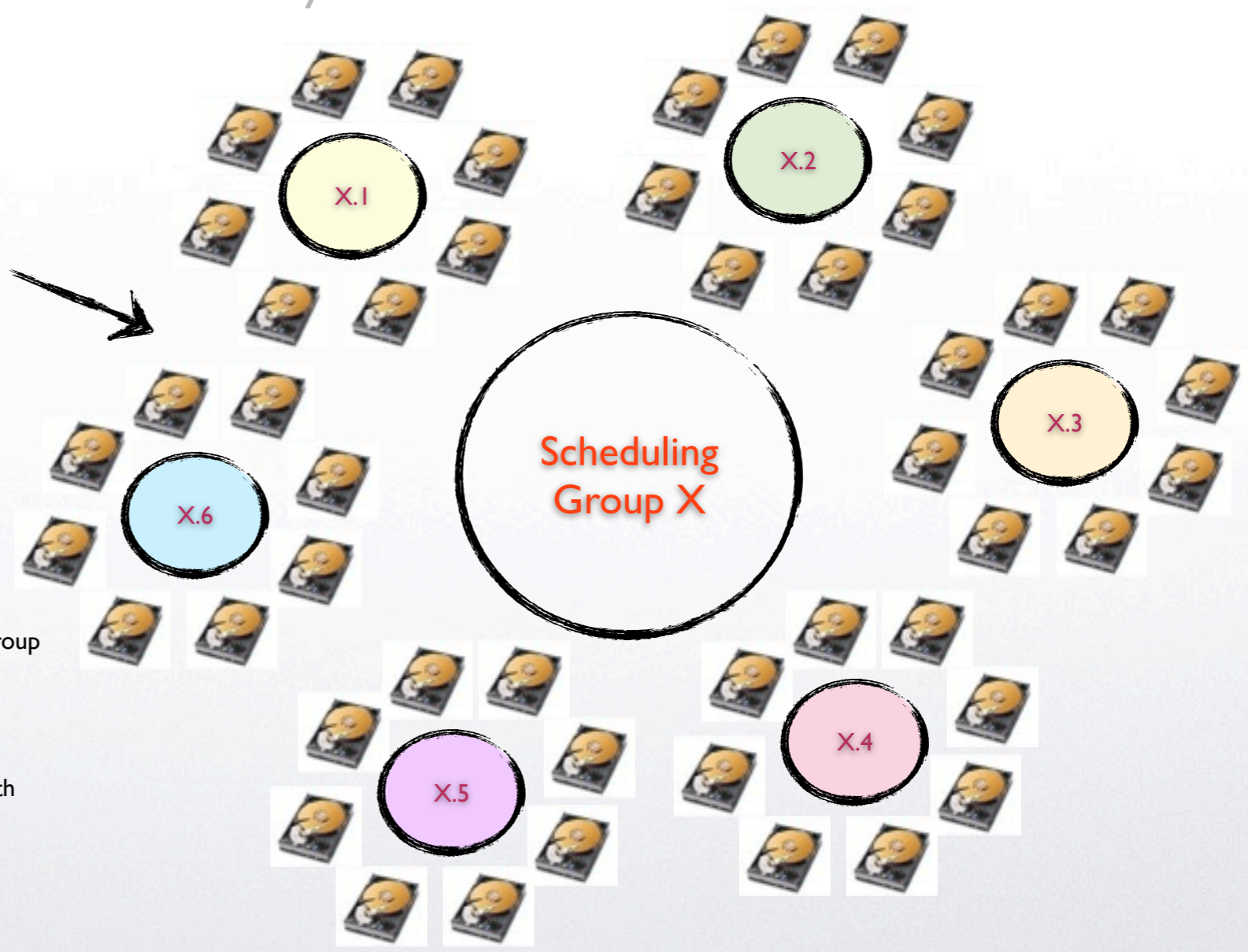
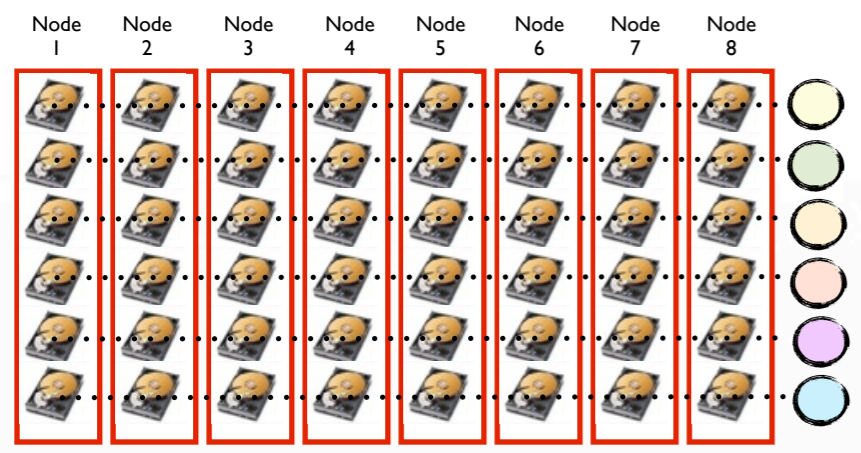
# Characteristics

- **Storage with single disks (JBODs-no RAID arrays)**
  - redundancy by s/w using cheap and unreliable h/w
- **Network RAID within disk groups**
  - scheduling (sub-)groups & round-robin rings
- **Online filesystem migration**
- **Tunable quality of service**
  - via redundancy parameters
- **Tradeoff in Scalability vs Latency**
  - namespace size, number of disks to manage



# Scheduling Groups

Redundancy over nodes



### Mapping of node/disks to **scheduling groups** and **scheduling subgroups**

- replicas are placed within one subgroup
- scheduling subgroup is changed **round-robin** with each new file placement
- the placement algorithm remembers round-robin pointers in each group for individual users and datasets

**Goal:** even distribution of files to maximize IOOPs and throughput

**IO Balancing:** each possible file placement or access is weighted with the current measured IO load on the corresponding disk

Version 0.1.0



# EOS Features 0.1.0 (I)

- **JBOD replica layout**
- User & Group **Quota** Nodes (quota attached to directory subtrees)
- **Disk Re-Balancing** (when disks are added)
- **Disk Draining** (when disks are to be removed)
- **File Pre-Allocation** (guarantees that a file can be written if the size is pre-defined)
- **File Checksums**
- **Block Checksums**
- Active Namespace Redirection on ENOENT and ENONET (file not found or file not available)
- directory based **ACL + E-GROUP support** (R,W & WO [no delete, no update])

Version 0.1.0



# EOS Features 0.1.0 (2)

- **Access interface** to ban, redirect and stall user
- **Error console** to follow errors of any server in installation
- File System Integrity Check (**FSCK**)
- Namespace & IO **statistic interface**
- **Virtual ID** Configuration (admin role, sudo permissions in the filesystem)
- **HA** daemon EOSHA for high-availability MGM master-slave failover
- Low-Level **FUSE** Implementation for shared mounts with krb5/x509 auth (eosd)
- Default **FUSE** Implementation for user private mounts with krb5/x509 auth (eosfsd)
- **File/Block-Checksum scanning** in defined intervals with disk-load-adaptive scan speed

Version 0.1.0



# EOS Shell

```
EOS Console [root://localhost] |/> help
access      Access Interface
attr        Attribute Interface
clear       Clear the terminal
cd          Change directory
chmod       Mode Interface
chown       Chown Interface
config      Configuration System
console     Run Error Console
debug       Set debug level
exit        Exit from EOS console
file        File Handling
fileinfo    File Information
find        Find files/directories
fs          File System configuration
fsck        File System Consistency Checking
fuse        Fuse Mounting
group       Group configuration
help        Display this text
io          IO Interface
license     Display Software License
ls          List a directory
mkdir       Create a directory
motd        Message of the day
node        Node configuration
ns          Namespace Interface
vid         Virtual ID System Configuration
pwd         Print working directory
quit        Exit from EOS console
quota       Quota System configuration
restart     Restart System
rmdir       Remove a directory
rm          Remove a file
role        Set the client role
rtlog       Get realtime log output from mgm & fst servers
silent      Toggle silent flag for stdout
space       Space configuration
test        Run performance test
timing      Toggle timing flag for execution time measurement
transfers   Transfer Interface
verify      Verify Interface
version     Verbose client/server version
whoami      Determine how we are mapped on server side
who         Statistics about connected users
?           Synonym for 'help'
.o         Exit from EOS console
```

## Interactive Shell with completion:

```
[root@eosdevsrv1 ~]# eos
=> selected user role ruid=<0> and group role rgid=<0>
#####
# Welcome to EOSDEV - have a nice day #
#####
EOS_INSTANCE=eosdev
EOS_SERVER_VERSION=0.1.0 EOS_SERVER_RELEASE=rc24
EOS_CLIENT_VERSION=0.1.0 EOS_CLIENT_RELEASE=rc24
EOS Console [root://localhost] |/>
```

## Non-Interactive Shell CMDs:

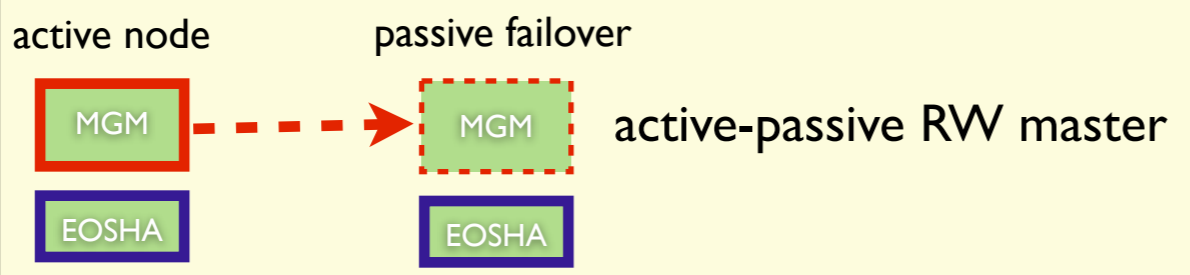
```
[root@eosdevsrv1 ~]# eos -b ns
# -----
# Namespace Statistic
# -----
ALL      Files                1099778
ALL      Directories           24006
# .....
ALL      File Changelog Size   1.83 GB
ALL      Dir Changelog Size    7.39 MB
# .....
ALL      avg. File Entry Size  1.67 kB
ALL      avg. Dir Entry Size   307.00 B
# -----
```

Version 0.1.0



# Namespace High Availability (HA) Instance Separation

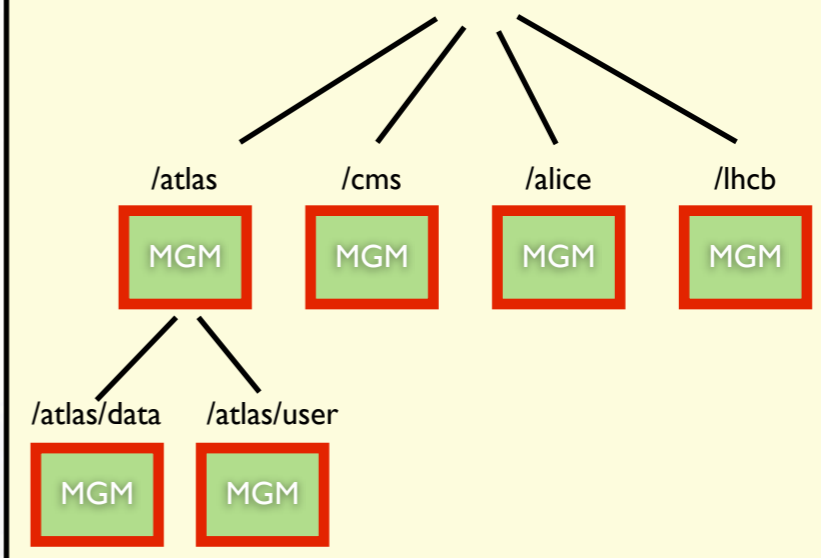
## HA



- CERN DNS loadbalancing alias is used to select RW master
- **EOSHA** daemon checkpoints the alias and start/stop the daemon
- Failover time 3-4 min.

*DNS solution is not optimal, evaluating alternatives like Linux HA*

## Instance Separation



For practical reasons we divide by experiment, further splitting is not (yet) needed for the expected scale

*Version 0.1.0*





# Production Instances



# Currently 4 Instances: DEV, PPS, ATLAS, CMS

## EOS ATLAS 1.9k disks - 3.8 PB

```
EOS Console [root://localhost] |/> space ls
```

#	type #	name #	groupsize #	groupmod	#N(fs)	#N(fs-rw)	#sum(usedbytes)	#sum(capacity)	#capacity(rw)	#nom.capacity	#quota	#balancing	#threshold
spaceview		default	24	24	1934	1858	2.75 P	3.81 P	3.71 P	3.00 P	on	off	100.00 G
spaceview		spare	24	24	43	42	1.98 G	83.85 T	83.84 T	0.00	off	off	0.00

## EOS CMS 1.2k disks - 2.3 PB

```
EOS Console [root://localhost] |/> space ls
```

#	type #	name #	groupsize #	groupmod	#N(fs)	#N(fs-rw)	#sum(usedbytes)	#sum(capacity)	#capacity(rw)	#nom.capacity	#quota	#balancing	#threshold
spaceview		default	20	24	1195	1180	1.36 P	2.38 P	2.36 P	2.30 P	on	off	500.00 G
spaceview		spare	12	24	2341	2252	27.35 G	4.63 P	4.49 P	0.00	off	off	50.00 G

```
bash-3.2$ eos -b root://eosatlas fuse mount $PWD/eos-atlas
OK
===> Mountpoint : /afs/cern.ch/user/a/apeters/eos-atlas
===> Fuse-Options : kernel_cache,attr_timeout=30,entry_timeout=30,max_readahead=131072,max_write=4194304,fsname=eosatlas root://eosatlas//eos/
===> xrootd ra : 4000000
===> xrootd cache : 16000000
bash-3.2$ eos -b root://eoscms fuse mount $PWD/eos-cms
OK
===> Mountpoint : /afs/cern.ch/user/a/apeters/eos-cms
===> Fuse-Options : kernel_cache,attr_timeout=30,entry_timeout=30,max_readahead=131072,max_write=4194304,fsname=eoscms root://eoscms//eos/
===> xrootd ra : 4000000
===> xrootd cache : 16000000
bash-3.2$ df | grep eos
eosatlas 3721425198048 2682298401436 1039126796612 73% /afs/cern.ch/user/a/apeters/eos-atlas
eoscms 2327583131584 1331544736592 996038394992 58% /afs/cern.ch/user/a/apeters/eos-cms
bash-3.2$ df -H | grep eos
eosatlas 3.9P 2.8P 1.1P 73% /afs/cern.ch/user/a/apeters/eos-atlas
eoscms 2.4P 1.4P 1.1P 58% /afs/cern.ch/user/a/apeters/eos-cms
```

EOS via FUSE  
(possible on lxplus at CERN)

Production Instances



# Experiment Migration Plans to EOS

## CMS

Pool	Current Size	Proposed Size	Proposed Date	Prerequisites	LFN Area	#files	Status
CMSCAF	1.7PB	2PB	June 20-27	Switch to xrootd access (28th March), <a href="#">PhEDEx</a> node T2_CH_CERN	/store/	992948	Done
CMSCAFUSER [*]	210TB	300TB	September	Data only under /store/caf/, CRAB stageout	/store/caf/	810051	
GRIDHOME [*]	50TB	100TB	September	Users need to update their CRAB config, redirect rule needed	/store/user/	13158	
CMSCAFT2 [*]	80TB	300TB	September	Redirect rule can be removed, CRAB stageout	/store/user/	22919	
DEFAULT	220TB	600TB	September	Dataset Popularity Service?	/store/	562143	
CMST3 [*]	420TB	420TB [1]	September	Users need to update their CRAB config, data only under /store/cmst3/, CRAB stageout	/store/cmst3/	1430045	
<a href="#">TOEXPRESS</a> [*]	200TB	200TB	November	Tier-0 supports stage-out to xrootd	/store/express/	1762192	
<a href="#">TOTEMP</a>	150TB	150TB	November	Tier-0 supports stage-out to xrootd	/store/t0temp/	373100	
<a href="#">TOSTREAMER</a> [2]	500TB	500TB	?	P5 transfer supports stage-out to xrootd	/store /tostreamer/	9774998	

## ATLAS

Already migrated

- atlasdata 1.5 PB
- atlascratch 50 TB

13th September

- atlascernngroupdisk
- atlascernuserdisk

Other instances are envisaged ...

Production Instances



# Default Configuration

```
EOS Console [root://localhost] |/> attr ls /eos/atlas
sys.forced.blockchecksum="crc32c"
sys.forced.blocksize="4k"
sys.forced.checksum="adler"
sys.forced.layout="replica"
sys.forced.nstripses="2"
```

- 4k block checksum calculation with SSE4 hardware CRC32C (no visible effect on RW performance on 10G box)  
 - Adler File Checksumming  
 - 2 Replicas

## ATLAS

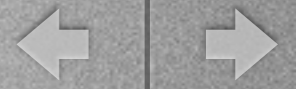
```
EOS Console [root://localhost] |/> quota ls
#
# ==> Quota Node: /eos/atlas/atlasdatadisk/
#
user      used bytes logi bytes used files aval bytes aval logib aval files filled[%] status
cstener  0.00 B    0.00 B    0.00 -    0.00 B    0.00 B    0.00 -    100.00  ignored
atlas00  2.67 PB   1.33 PB   6.15 M-   0.00 B    0.00 B    0.00 -    100.00  ignored
atlassg  1.55 kB   774.00 B  1.00 -    0.00 B    0.00 B    0.00 -    100.00  ignored
# .....
group    used bytes logi bytes used files aval bytes aval logib aval files filled[%] status
zp       2.67 PB   1.33 PB   6.15 M-   3.00 PB   1.50 PB   8.00 M-   88.99   ok
#
# ==> Summary
user      used bytes logi bytes used files aval bytes aval logib aval files filled[%] status
ALL      2.67 PB   1.33 PB   6.15 M-   0.00 B    0.00 B    0.00 -    100.00  ignored
group    used bytes logi bytes used files aval bytes aval logib aval files filled[%] status
ALL      2.67 PB   1.33 PB   6.15 M-   3.00 PB   1.50 PB   8.00 M-   88.99   ok
```

+ individual user quota for ~ 450 users on ATLAS

## CMS

... uses more fine grained quota on II quota nodes  
 + individual user quota for ~ 800 users on CMS

Production Instances



# EOS Operations




# Availability

## Availability is monitored via SLS service at CERN

**EOS data storage service**

---

EOS



availability:   
(more)

percentage: 100%

status: **available**

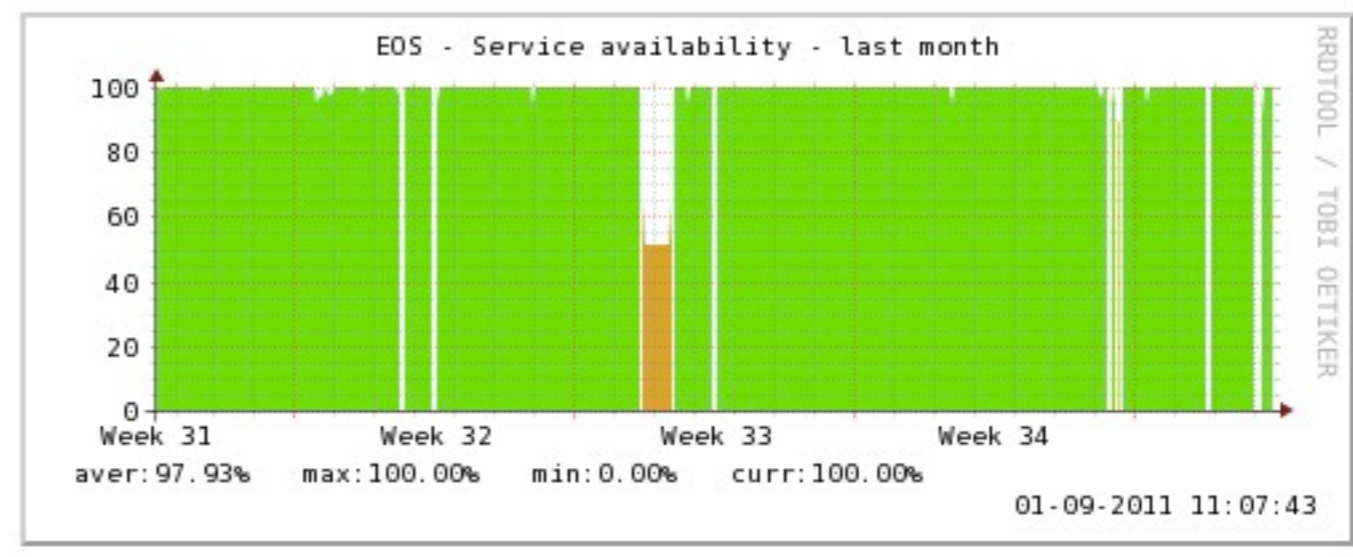
---

this service consists of:

-  EOSATLAS - EOS storage service for ATLAS
-  EOSCMS - EOS storage service for CMS
-  EOSPPS

Measured via xrdcp + lcg-cp probe (up-,download,deletion)

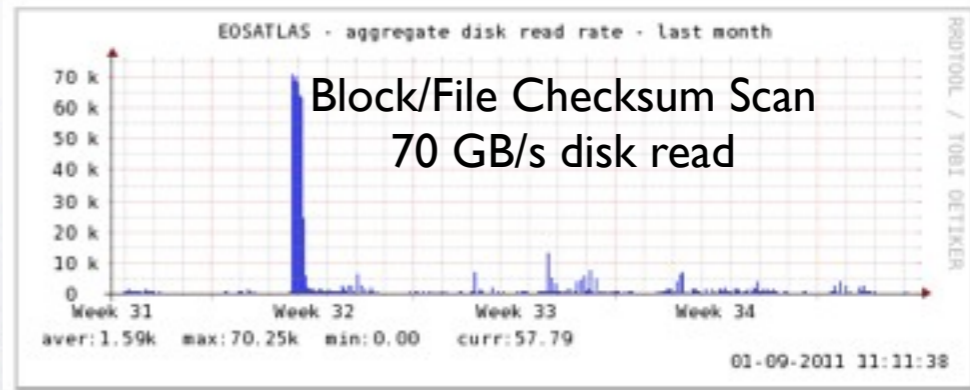
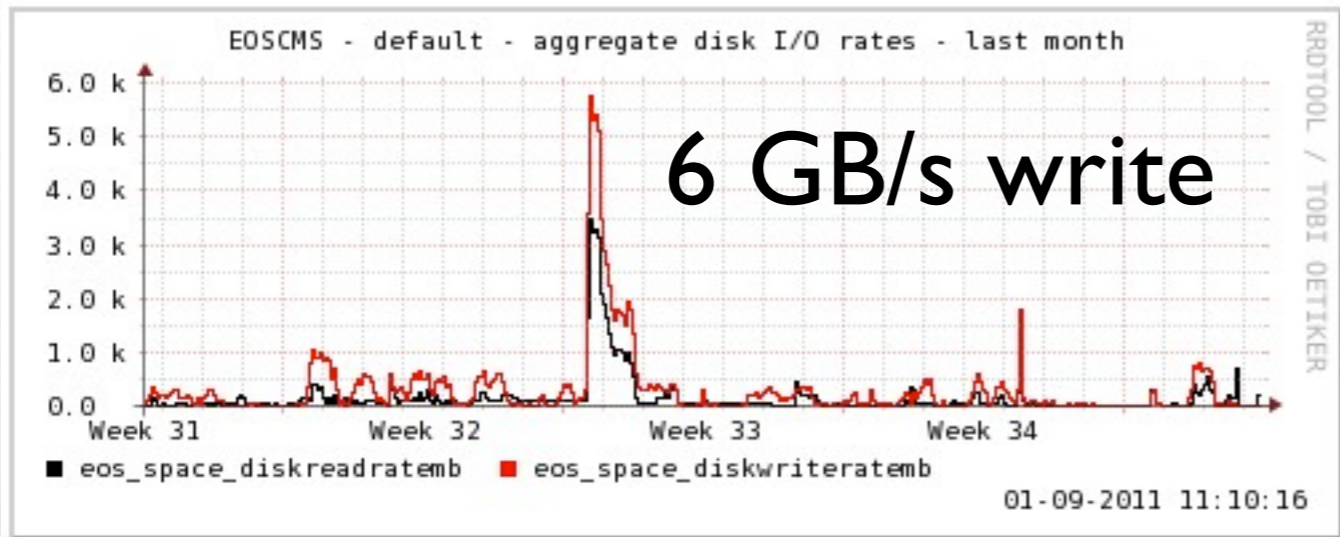
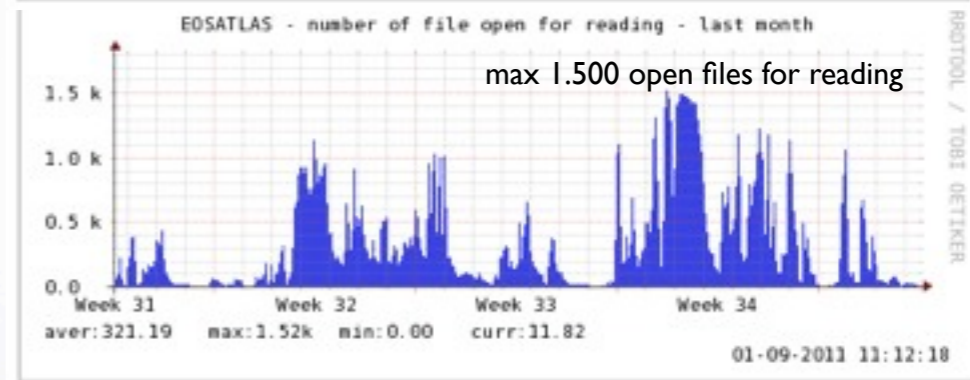
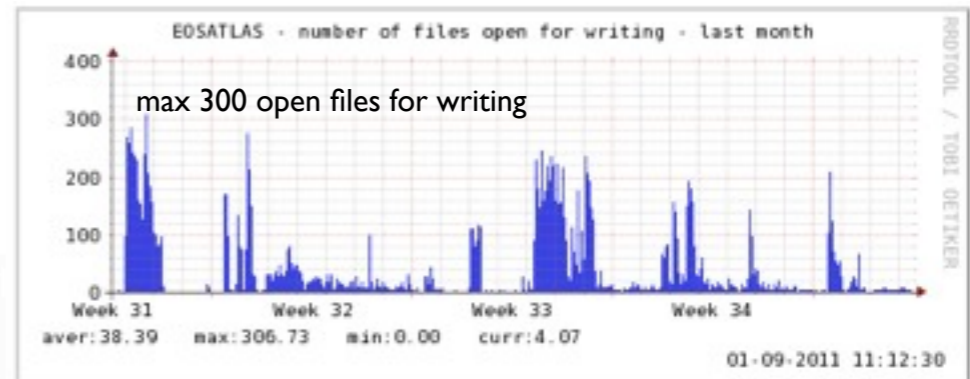
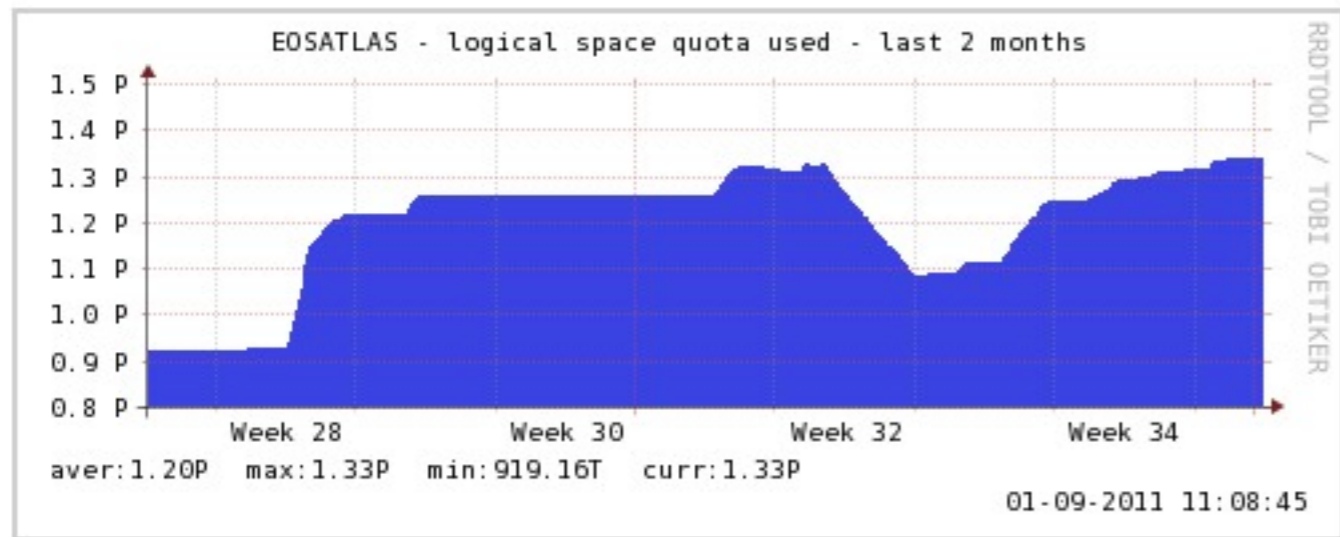
Including scheduled interventions and probe misconfiguration (orange)  
**98%** available during last month



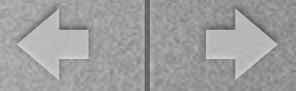
**EOS Operations**



# Usage



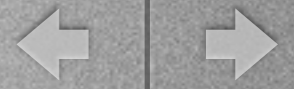
EOS Operations








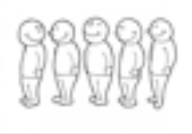
# Error Handling

- **Challenge:** run without piquet, best effort support
  - Failures don't require immediate human interventions
    - MGM failover via **EOSHA**
    - Disks drain automatically triggered by IO or pattern scrubbing errors after a configurable grace period
      - drain time on production instance < 1h for 2 TB disk (10-20 disks per scheduling group)
  - Sysadmin team replaces disks 'asynchronously' using admin tools to remove and re-add filesystems to EOS
  - Procedure & software support is still undergoing refinement/fixing





# Roadmap

- **EOS 0.1.0** Release candidate used in EOSCMS/EOSATLAS (still bug fixing)
- **EOS 0.2.0** in late autumn
  - **DPR/ZFEC** - Dual Parity Raid Layout Driver (like file-level Raid-6 over hosts) + ZFEC Driver (Reed-Solomon) 
  - **DPR/ZFEC** check & recovery tool: 
  - **Atomicity** for multiple writers on a file 
  - **Directory Cache** for low-level FUSE mount 
  - **OSX/Linux Client bundle** for User EOS mounting (krb5 or GSI) 
  - **cmsd** plugin for global xrootd subscription  in the queue

EOS Roadmap



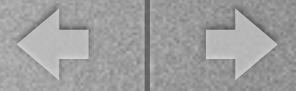
# Summary & Outlook

- Two production instances running
  - result of very good cooperation with experiments
- Expand usage & gain more experience
- Move from **rapid development** done during last 15 month to **reliable production mode**
  - mutual agreement of development, operations team & experiments

**Final remark:** will not attempt production deployment outside CERN before main goals have been achieved there



Thank you  
for your attention!



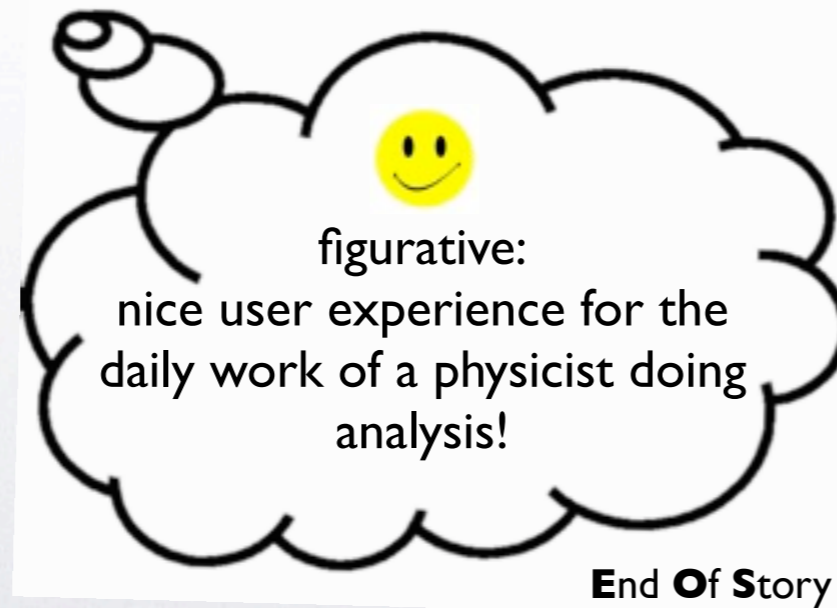
What does  
EOS stand for?



# EOS



- In [Greek mythology](#), **Eos** ( [/ˈiːɒs/](#); [Greek](#): Ἠώς, or "Εως "dawn", pronounced [\[ɛːɔ̃ːs\]](#) or [\[éɔːs\]](#)) is the [Titan goddess](#) of the dawn, who rose from her home at the edge of [Oceanus](#), the ocean that surrounds the world, to herald her brother [Helios](#), the Sun.
- The dawn goddess, Eos with "rosy fingers" opened the gates of heaven<sup>[2]</sup> so that [Helios](#), her brother, could ride his chariot across the sky every day





# Appendix



Spaces segment the pool hardware

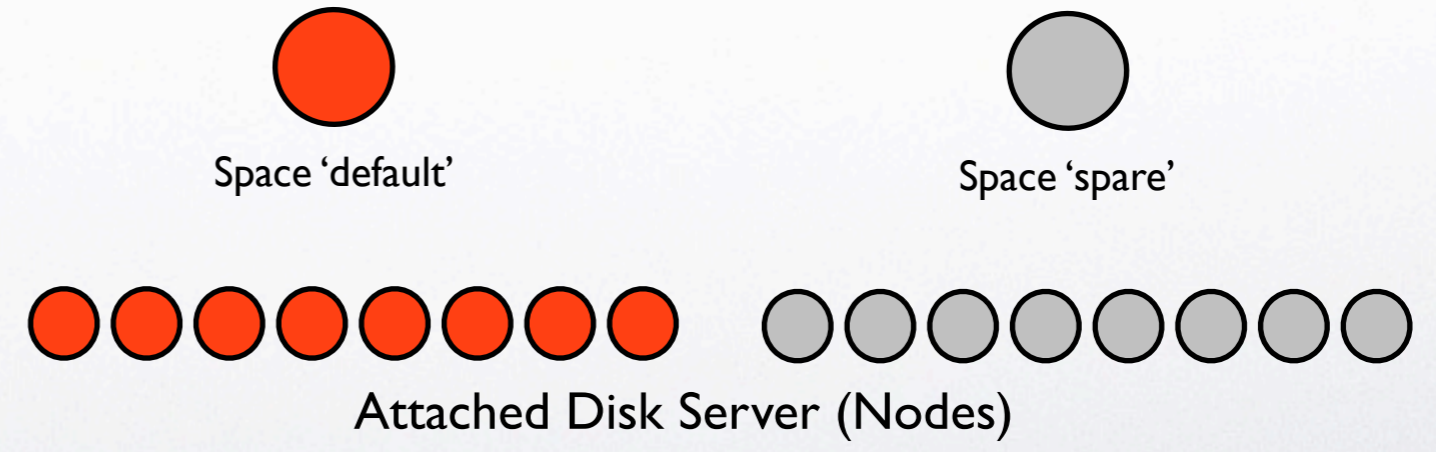
# Space View

```
EOS Console [root://localhost] |/> space ls default
#-----
#   type #          name #  groupsize #  groupmod #N(fs) #N(fs-rw) #sum(usedbytes) #sum(capacity) #capacity(rw) #nom.capacity #quota #balancing #threshold
#-----
spaceview          default          24          24    1933    1858          2.75 P          3.81 P          3.71 P          3.00 P    on    off    100.00 G

EOS Console [root://localhost] |/> space ls --io default
#-----
#   name # diskload # diskr-MB/s # diskw-MB/s #eth-MiB/s # ethi-MiB # etho-MiB #ropen #wopen #used-bytes # max-bytes # used-files # max-files
#-----
default      0.01      131.00      17.00      9475      70        87        52      4      2.75 PB    3.81 PB    13.44 M    182.19 G
```

## Information about

- number of disks
- used/max space
- used/max inodes
- load of the disks (0 - 1.0)
- IO rates for disk + net
- Open files for read/write



Production Instances



# FileSystem View

EOS Console [root://localhost] |/> fs ls -l

#	host #	port #	id #	uuid #	path #	schedgroup #	headroom #	boot #	configstatus #	drain #	active#	scaninterval
1	lxfsrg03a03.cern.ch	1095	1	05d844ef-351f-43dd-a4ff-2d17a9dedff2	/data01	default.76	25.00 G	booted	rw	nodrain	online	2592000
2	lxfsrg03a03.cern.ch	1095	2	745c8f2a-985b-43c2-9d0d-7fc90b3740ea	/data02	default.81	25.00 G	booted	rw	nodrain	online	2592000
3	lxfsrg03a03.cern.ch	1095	3	7bae66f5-45a5-47b9-8d87-149e1fb5a990	/data03	default.18	25.00 G	booted	rw	nodrain	online	2592000
4	lxfsrg03a03.cern.ch	1095	4	f48c38af-ddc9-4ace-86a4-9d5214585fa3	/data04	default.78	25.00 G	booted	rw	nodrain	online	2592000
5	lxfsrg03a03.cern.ch	1095	5	3d71e12e-e61a-4acd-af17-a255cee64372	/data05	default.97	25.00 G	booted	rw	nodrain	online	2592000
6	lxfsrg03a03.cern.ch	1095	6	d2d41dc3-912b-448f-8d0a-fa392e754886	/data06	default.22	25.00 G	booted	rw	nodrain	online	2592000
7	lxfsrg03a03.cern.ch	1095	7	81fa74a2-959f-45fb-870a-afe7026eca50	/data07	default.62	25.00 G	booted	rw	nodrain	online	2592000
8	lxfsrg03a03.cern.ch	1095	8	647ec450-783e-4a19-a12f-53cde5ea3640	/data08	default.4	25.00 G	booted	rw	nodrain	online	2592000
9	lxfsrg03a03.cern.ch	1095	9	5fe7cd1a-2412-4db6-beba-463ec4906877	/data09	default.40	25.00 G	booted	rw	nodrain	online	2592000
10	lxfsrg03a03.cern.ch	1095	10	2087e128-d5ae-4c83-bb2c-8ad04a464694	/data10	default.94	25.00 G	booted	rw	nodrain	online	2592000
11	lxfsrg03a03.cern.ch	1095	11	b8636372-57fb-46df-a030-191f013a9303	/data11	default.47	25.00 G	booted	rw	nodrain	online	2592000
12	lxfsrg03a03.cern.ch	1095	12	3394c803-fe6d-483e-8ef6-d1bdf559a743	/data12	default.53	25.00 G	booted	rw	nodrain	online	2592000
13	lxfsrg03a03.cern.ch	1095	13	d21f6321-fd2c-4a22-8df8-33906676a6ff	/data13	default.23	25.00 G	booted	rw	nodrain	online	2592000
14	lxfsrg03a03.cern.ch	1095	14	22175df5-eac2-4229-a87d-6215b2d9c3c5	/data14	default.5	25.00 G	booted	rw	nodrain	online	2592000
15	lxfsrg03a03.cern.ch	1095	15	a74dff9f-723a-4d12-ad61-508696186b8a	/data15	default.99	25.00 G	booted	rw	nodrain	online	2592000
16	lxfsrg03a03.cern.ch	1095	16	7c02f055-cf45-4b52-9071-63e40041cca2	/data16	default.87	25.00 G	booted	rw	nodrain	online	2592000

EOS Console [root://localhost] |/> fs ls -e

#	host #	id #	path #	boot #	configstatus #	drain #...	#errmsg
63	lxfsrg03a06.cern.ch	63	/data21	bootfailure	empty	drained	5 cannot have <rw> access
156	lxfsrg05a06.cern.ch	156	/data07	bootfailure	empty	drained	5 cannot have <rw> access
298	lxfsrc56a01.cern.ch	298	/data13	bootfailure	empty	drained	5 cannot have <rw> access
665	lxfsrd63a01.cern.ch	665	/data08	bootfailure	empty	drained	5 cannot have <rw> access
674	lxfsrd63a01.cern.ch	674	/data17	bootfailure	empty	drained	5 cannot have <rw> access
679	lxfsrd63a01.cern.ch	679	/data22	bootfailure	empty	drained	5 cannot have <rw> access
691	lxfsrd63a03.cern.ch	691	/data12	bootfailure	empty	drained	5 cannot have <rw> access
817	lxfsrd63a08.cern.ch	817	/data02	bootfailure	empty	drained	5 cannot have <rw> access
1041	lxfsre01a08.cern.ch	1041	/data18	bootfailure	empty	drained	5 cannot have <rw> access
1407	lxfsrg09a06.cern.ch	1407	/data15	bootfailure	empty	drained	5 cannot have <rw> access
1513	lxfsrg11a03.cern.ch	1513	/data12	bootfailure	empty	drained	5 cannot have <rw> access
1652	lxfsrg13a01.cern.ch	1652	/data01	bootfailure	empty	drained	5 cannot have <rw> access
1742	lxfsrg13a05.cern.ch	1742	/data05	bootfailure	empty	drained	5 cannot have <rw> access
1798	lxfsrg13a07.cern.ch	1798	/data10	bootfailure	empty	drained	14 cannot write the filesystem label (fsid+uuid) - please check filesystem state/permissions

Provides per File System Parameters

Production Instances





# File System View - Disk Failures

EOS Console [root://localhost] |> fs ls -d

#	host (#...)	#	id #	path #	drain #	progress #	files #	lost-files #	bytes-left	#sched-files	#sched-bytes	# graceperiod	timeleft	#retry
lxfsrg03a03.cern.ch (1095)	20	/data20	drained	100	0.00 B	0	0.00	0.00	0.00	0	0	0	0	0
lxfsrg03a06.cern.ch (1095)	63	/data21	drained	100	0.00 B	0	0.00	0.00	0.00	0	0	0	0	0
lxfsrg05a02.cern.ch (1095)	110	/data04	drained	100	0.00 B	0	0.00	0.00	0.00	0	0	0	0	0
lxfsrg05a04.cern.ch (1095)	146	/data19	drained	100	0.00 B	0	0.00	0.00	0.00	0	0	0	0	0
lxfsrg05a06.cern.ch (1095)	156	/data07	drained	100	0.00 B	0	0.00	0.00	0.00	0	0	0	0	0
lxfsra24a01.cern.ch (1095)	275	/data12	drained	100	0.00 B	0	0.00	0.00	0.00	0	0	0	0	0
lxfsrc56a01.cern.ch (1095)	298	/data13	drained	100	0.00 B	0	0.00	0.00	0.00	0	0	0	0	0
lxfsrc56a01.cern.ch (1095)	304	/data19	drained	100	0.00 B	0	0.00	0.00	0.00	0	0	0	0	0
lxfsrd63a01.cern.ch (1095)	661	/data04	drained	100	0.00 B	0	0.00	0.00	0.00	0	0	0	0	0
lxfsrd63a01.cern.ch (1095)	665	/data08	drained	100	0.00 B	0	0.00	0.00	0.00	0	0	0	0	0
lxfsrd63a01.cern.ch (1095)	674	/data17	drained	100	0.00 B	0	0.00	0.00	0.00	0	0	0	0	0
lxfsrd63a01.cern.ch (1095)	679	/data22	drained	100	0.00 B	0	0.00	0.00	0.00	0	0	0	0	0
lxfsrd63a03.cern.ch (1095)	691	/data12	drained	100	0.00 B	0	0.00	0.00	0.00	0	0	0	0	0
lxfsrd63a08.cern.ch (1095)	817	/data02	drained	100	0.00 B	0	0.00	0.00	0.00	0	0	0	0	0
lxfsre01a08.cern.ch (1095)	1041	/data18	drained	100	0.00 B	0	0.00	0.00	0.00	0	0	0	0	0
lxfsrg09a03.cern.ch (1095)	1357	/data08	drained	100	0.00 B	0	0.00	0.00	0.00	0	0	0	0	0
lxfsrg09a06.cern.ch (1095)	1407	/data15	drained	100	0.00 B	0	0.00	0.00	0.00	0	0	0	0	0
lxfsrg11a03.cern.ch (1095)	1506	/data06	drained	100	0.00 B	0	0.00	0.00	0.00	0	0	0	0	0
lxfsrg11a03.cern.ch (1095)	1508	/data08	drained	100	0.00 B	0	0.00	0.00	0.00	0	0	0	0	0
lxfsrg11a03.cern.ch (1095)	1513	/data12	drained	100	0.00 B	0	0.00	0.00	0.00	0	0	0	0	0
lxfsrg11a05.cern.ch (1095)	1550	/data06	drained	100	0.00 B	0	0.00	0.00	0.00	0	0	0	0	0
lxfsrg11a05.cern.ch (1095)	1562	/data18	drained	100	0.00 B	0	0.00	0.00	0.00	0	0	0	0	0
lxfsrg13a01.cern.ch (1095)	1652	/data01	drained	100	0.00 B	0	0.00	0.00	0.00	0	0	0	0	0
lxfsrg13a05.cern.ch (1095)	1742	/data05	drained	100	0.00 B	0	0.00	0.00	0.00	0	0	0	0	0
lxfsrg13a06.cern.ch (1095)	1769	/data10	drained	100	0.00 B	0	0.00	0.00	0.00	0	0	0	0	0
lxfsrg13a06.cern.ch (1095)	1771	/data12	drained	100	0.00 B	0	0.00	0.00	0.00	0	0	0	0	0
lxfsrg13a06.cern.ch (1095)	1777	/data18	drained	100	0.00 B	0	0.00	0.00	0.00	0	0	0	0	0
lxfsrg13a07.cern.ch (1095)	1798	/data10	drained	100	0.00 B	0	0.00	0.00	0.00	0	0	0	0	0
lxfsrg15a01.cern.ch (1095)	1827	/data02	drained	100	0.00 B	0	0.00	0.00	0.00	0	0	0	0	0
lxfsrg15a02.cern.ch (1095)	1853	/data06	drained	100	0.00 B	0	0.00	0.00	0.00	0	0	0	0	0

Example of EOS ATLAS: 30 drained disk

Production Instances



# Group View Balancing

```

EOS Console [root://localhost] |/> space ls
#-----#
#   type #           name #  groupsize #  groupmod #N(fs) #N(fs-rw) #sum(usedbytes) #sum(capacity) #capacity(rw) #nom.capacity #quota #balancing #threshold
#-----#
spaceview          default          10           22    216      191      59.25 T      431.17 T      381.27 T      0.00  off      on      2.00 G

EOS Console [root://localhost] |/> group ls
#-----#
#   type #           name #  status #nofs #dev( usedbytes) #avg( usedbytes) #sig( usedbytes) #balancing #  queued
#-----#
groupview default.0      on    10      731.18 MB      300.55 GB      381.47 MB idle      0.00
groupview default.1      on    10      547.34 MB      297.12 GB      302.35 MB idle      0.00
groupview default.10     on    10      666.53 MB      300.98 GB      278.53 MB idle      0.00
groupview default.11     on    10      389.12 MB      296.82 GB      174.64 MB idle      0.00
groupview default.12     on    10      780.26 MB      300.36 GB      315.35 MB idle      0.00
groupview default.13     on    10      1.32 GB        271.19 GB      775.25 MB idle      0.00
groupview default.14     on    10      1.37 GB        295.36 GB      916.26 MB idle      0.00
groupview default.15     on    10      1.52 GB        271.80 GB      738.32 MB idle      0.00
groupview default.16     on    10      1.02 GB        271.16 GB      691.22 MB idle      0.00
groupview default.17     on    10      980.83 MB      266.40 GB      455.36 MB idle      0.00
groupview default.18     on    10      996.05 MB      271.59 GB      414.95 MB idle      0.00
groupview default.19     on    10      1.00 GB        269.97 GB      510.72 MB idle      0.00
groupview default.2      on    10      1.88 GB        292.48 GB      803.65 MB idle      0.00
groupview default.20     on    10      1.30 GB        295.01 GB      641.58 MB idle      0.00
groupview default.21     on     6      1.02 GB        573.09 GB      646.13 MB idle      0.00
groupview default.3      on    10      1.80 GB        302.61 GB      730.80 MB idle      0.00
groupview default.4      on    10      1.46 GB        296.28 GB      860.06 MB idle      0.00
groupview default.5      on    10      693.93 MB      306.64 GB      276.05 MB idle      0.00
groupview default.6      on    10      1.52 GB        299.73 GB      770.94 MB idle      0.00
groupview default.7      on    10      391.40 MB      296.08 GB      174.47 MB idle      0.00
groupview default.8      on    10      385.43 MB      300.46 GB      154.53 MB idle      0.00
groupview default.9      on    10      1.67 GB        296.09 GB      985.14 MB idle      0.00

```

Example of EOS DEV instance: 22 scheduling groups to balance

Production Instances



# Namespace + IO Statistics

## EOSATLAS

```
EOS Console [root://localhost] |> ns stat
# -----
# Namespace Statistic
# -----
ALL      Files                6682009
ALL      Directories          128295
# -----
ALL      File Changelog Size  1.16 GB
ALL      Dir Changelog Size   38.55 MB
# -----
ALL      avg. File Entry Size 173.00 B
ALL      avg. Dir Entry Size  300.00 B
# -----
```

## EOSCMS

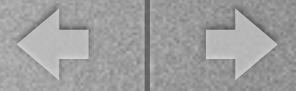
```
EOS Console [root://localhost] |> ns stat
# -----
# Namespace Statistic
# -----
ALL      Files                647129
ALL      Directories          112063
# -----
ALL      File Changelog Size  293.05 MB
ALL      Dir Changelog Size   40.87 MB
# -----
ALL      avg. File Entry Size 452.00 B
ALL      avg. Dir Entry Size  364.00 B
# -----
```

## EOSCMS

```
EOS Console [root://localhost] |> io stat
# -----
who      io value              sum      1min    5min    1h      24h
# -----
ALL      bytes_read            1.18 P  211.20 M  2.03 G  229.12 G  1.22 T
ALL      bytes_rseek          6.56 E  47.43 T  448.71 T  14.73 P  93.30 P
ALL      bytes_written        1.92 P   0.00   1.98 G   5.64 G   5.68 G
ALL      bytes_wseek          577.11 P 0.00   33.80 G  121.54 G  121.54 G
ALL      disk_time_read       18.30 G  30.20 k  221.25 k  6.93 M   58.95 M
ALL      disk_time_write      28.01 G   0.00   16.73 k  45.65 k  46.32 k
ALL      read_calls           21.94 G  7.69 k  84.34 k  13.12 M  54.96 M
ALL      write_calls           5.64 G   0.00   11.16 k  35.19 k  35.26 k
```

```
# -----
# top IO list by user name: bytes_read
# -----
[ bytes_read ] 1. cmsprod 529.49 T
[ bytes_read ] 2. relval 345.98 T
[ bytes_read ] 3. bin 183.93 T
[ bytes_read ] 4. venturia 41.14 T
[ bytes_read ] 5. mgrassi 13.29 T
[ bytes_read ] 6. aysen 11.64 T
[ bytes_read ] 7. obertino 7.48 T
[ bytes_read ] 8. hkseo 6.77 T
[ bytes_read ] 9. jkarancs 6.00 T
[ bytes_read ] 10. taroni 3.65 T
```

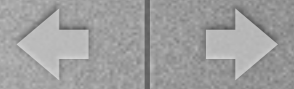
Production Instances



# Filesystem Check

- **fsck** tool collects with  $n$  parallel threads all meta data from all FSTs and creates a filesystem report
- Example: checks 25k Files/s
  - 2 Mio checked in 80s - depends on the load on the pool
- CLI to issue repair operations on the file system

Version 0.1.0



# Filesystem Check

```
EOS Console [root://localhost] |/> fsck
usage: fsck stat                               : print status of consistency check
      fsck enable [#threads]                  : enable fsck [with #threads threads]
      fsck disable                             : disable fsck
      fsck report [-h] [-g] [-m] [-a] [-i] [-l] [--error <tag>] : report consistency check results
al counters

      -m : select monitoring output format
      -a : break down statistics per filesystem
      -i : print concerned file ids
      -l : print concerned logical names
      --error <tag> : select only errors with name <tag> in the printout
                    : you get the names by doing 'fsck report -g'
      -h : print help explaining the individual tags!

      fsck repair --checksum                   : issues a 'verify' operation on all files with checksum errors
      fsck repair --unlink-unregistered       : unlink replicas which are not connected/registered to their logical name
      fsck repair --unlink-orphans           : unlink replicas which don't belong to any logical name
      fsck repair --adjust-replicas          : try to fix all replica inconsistencies
      fsck repair --drop-missing-replicas    : just drop replicas from the namespace if they cannot be found on disk

EOS Console [root://localhost] |/> █
```

Version 0.1.0



# Filesystem Check

```
EOS Console [root://localhost] |/> fsck report
ALL      totalfiles                2194425
ALL      diff_mgm_disk_size             0
ALL      diff_fst_disk_fmd_size        0
ALL      diff_mgm_disk_checksum        0
ALL      diff_fst_disk_fmd_checksum    7
ALL      diff_file_checksum_scan      7
ALL      diff_block_checksum_scan     0
ALL      scanned_files                2194422
ALL      not_scanned_files            3
ALL      replica_not_registered       0
ALL      replica_orphaned             0
ALL      diff_replica_layout          142
ALL      replica_offline              110
ALL      file_offline                 91
ALL      replica_missing              0
```

## Report Output

```
EOS Console [root://localhost] |/> fsck report -a -l --error file_offline
45      file_offline                4
lfn=/eos/dev/2rep/sub4/lxb8957.cern.ch_42/0/7/30.root e=file_offline
lfn=/eos/dev/2rep/sub4/lxb8954.cern.ch_20/0/7/73.root e=file_offline
lfn=/eos/dev/2rep/sub4/lxb8957.cern.ch_27/0/7/56.root e=file_offline
lfn=/eos/dev/2rep/sub4/lxb8955.cern.ch_24/0/7/61.root e=file_offline
47      file_offline                1
lfn=/eos/dev/2rep/sub3/lxb8954.cern.ch_8/0/4/22.root e=file_offline
51      file_offline                2
lfn=/eos/dev/2rep/sub3/lxb8954.cern.ch_40/0/0/45.root e=file_offline
lfn=/eos/dev/2rep/sub3/lxb8959.cern.ch_40/0/0/50.root e=file_offline
```

## Tracking Files with Errors

Version 0.1.0