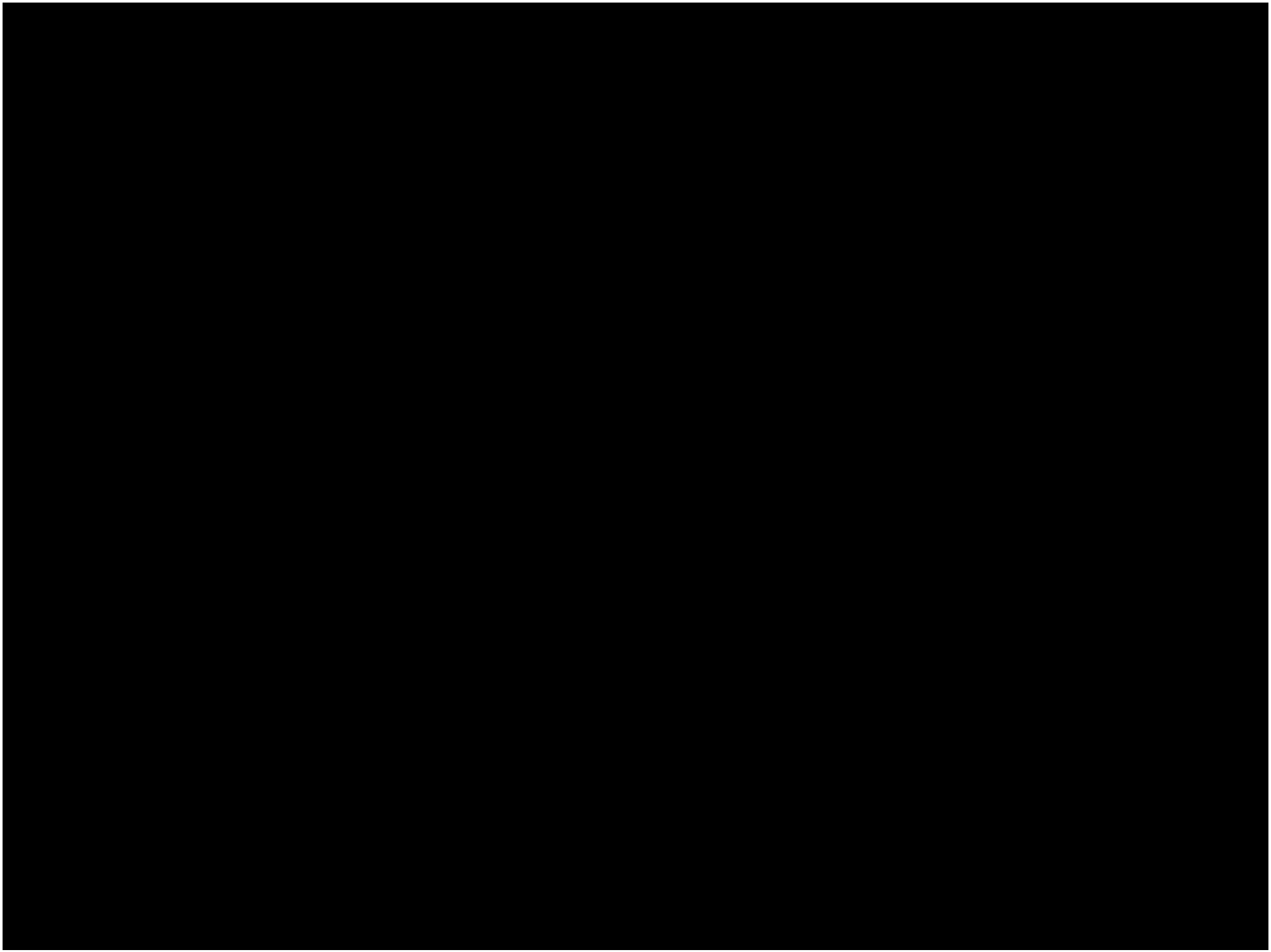


EOS

Features & Site Support

Andreas Peters
CERN IT-DSS

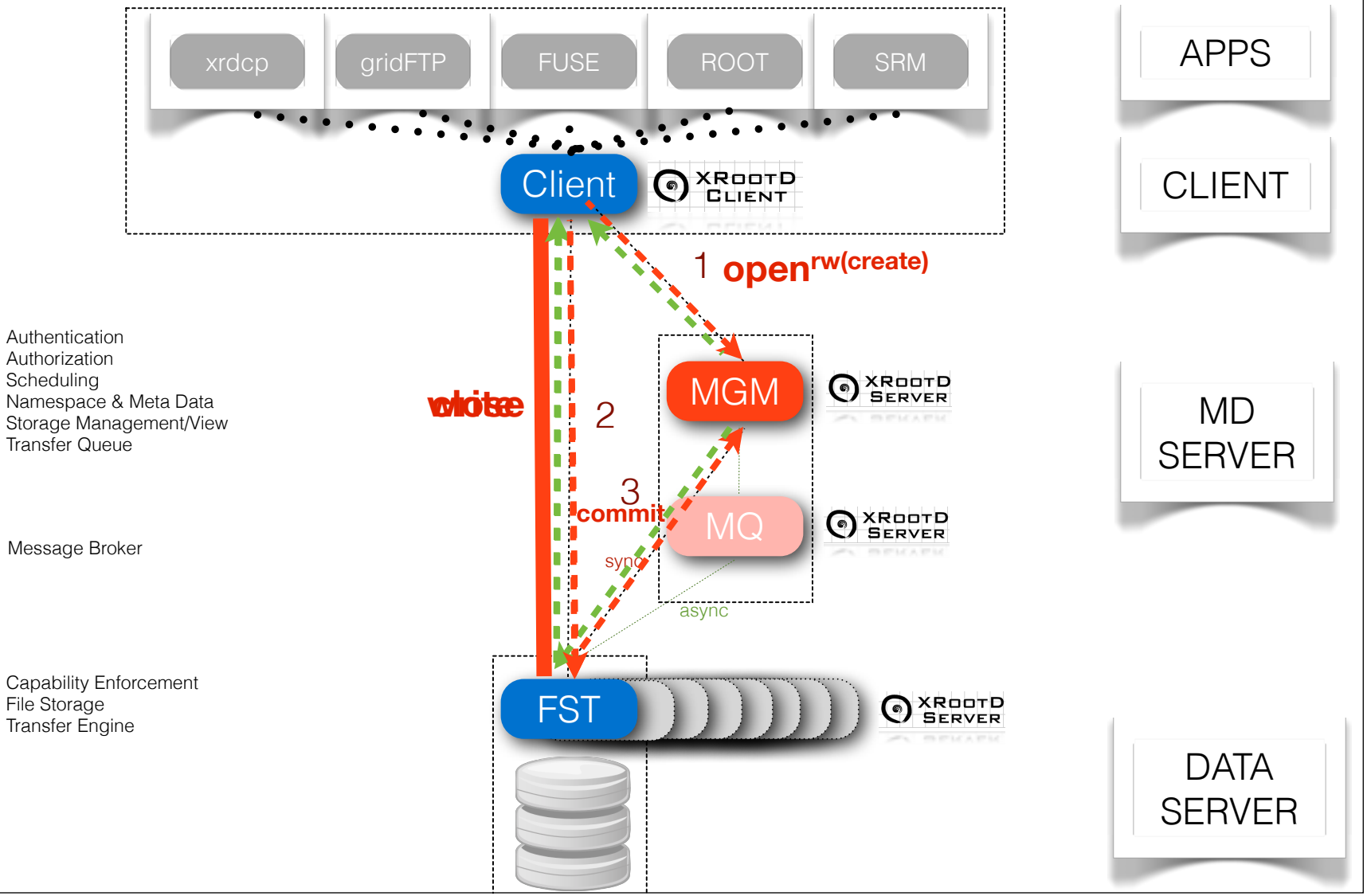
ALICE T1/2 Workshop 05.06.2013

Disclaimer

This presentation can spotlight only
a subset of many interesting topics and details.

Introduction

Base Architecture



Who Is Who

... is this a one man show ?

DSS Group

Alberto Pace

DT Section

Dirk Düllmann

FDO Section

Massimo Lamanna

EOS Project

approx. 2- 2.5 FTE

LJ	EAS	GA	AJP
Lukasz	Elvin	Geoffrey	Andreas
Namespace XRootD Release XRootD Client XrdCl		DSI/**gridFTP** Instr. **Mutex** LevelDB **DBmap**	
	XrdCl **Integration** **RAIN** **FUSE** **FAX**		Project Leader **Core** **Ops**

Development

approx. 1.3-1.5 FTE

ML	LM	JI	XV	BL
Massimo	Luca	Jan	Xavier	Belinda
Ops	**Ops** **Cockpit**	**Ops** **Puppetizer** **Lemon/SLS** Dev **Mentoring**	**Ops** **Puppetizer**	**Ops**

Operations

EOS@FNAL (Catalin D.)

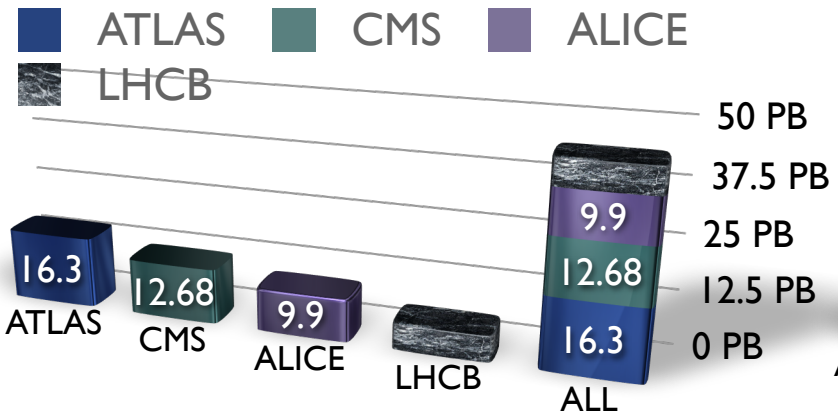
SysAdmin Team / Help Desk

IT-ES + external experiment contacts

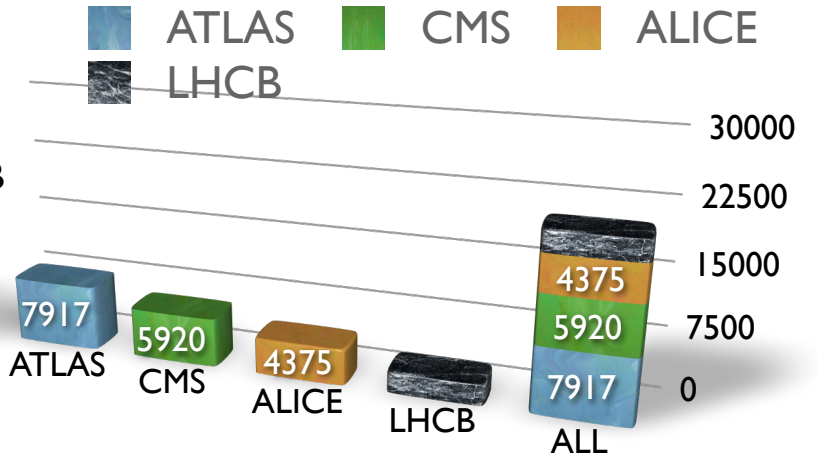
User Community in ATLAS,ALICE,CMS,LHCB,AMS,COMPASS

Service Today

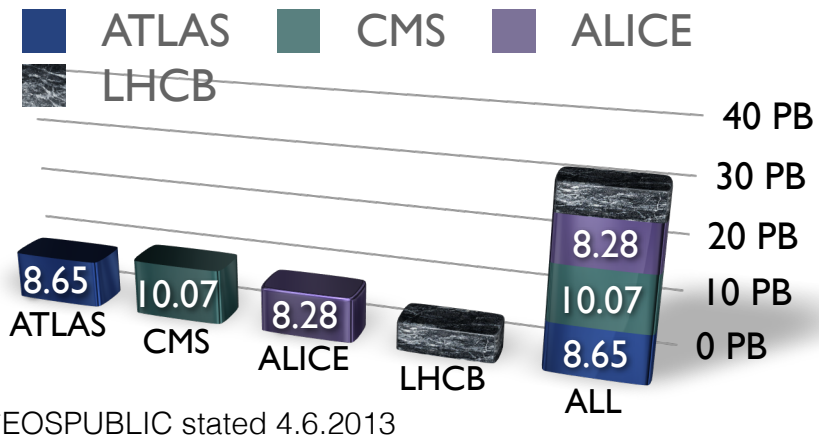
Raw Space **44.8 PB**



Harddisks **20.7k**

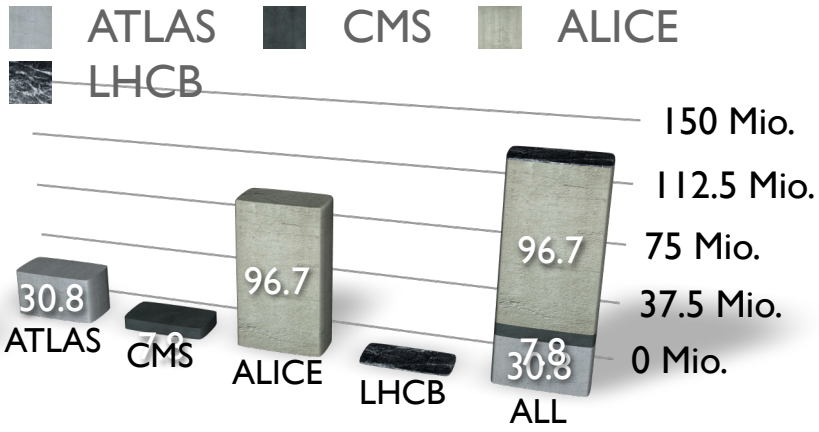


Used Space **32.1 PB**

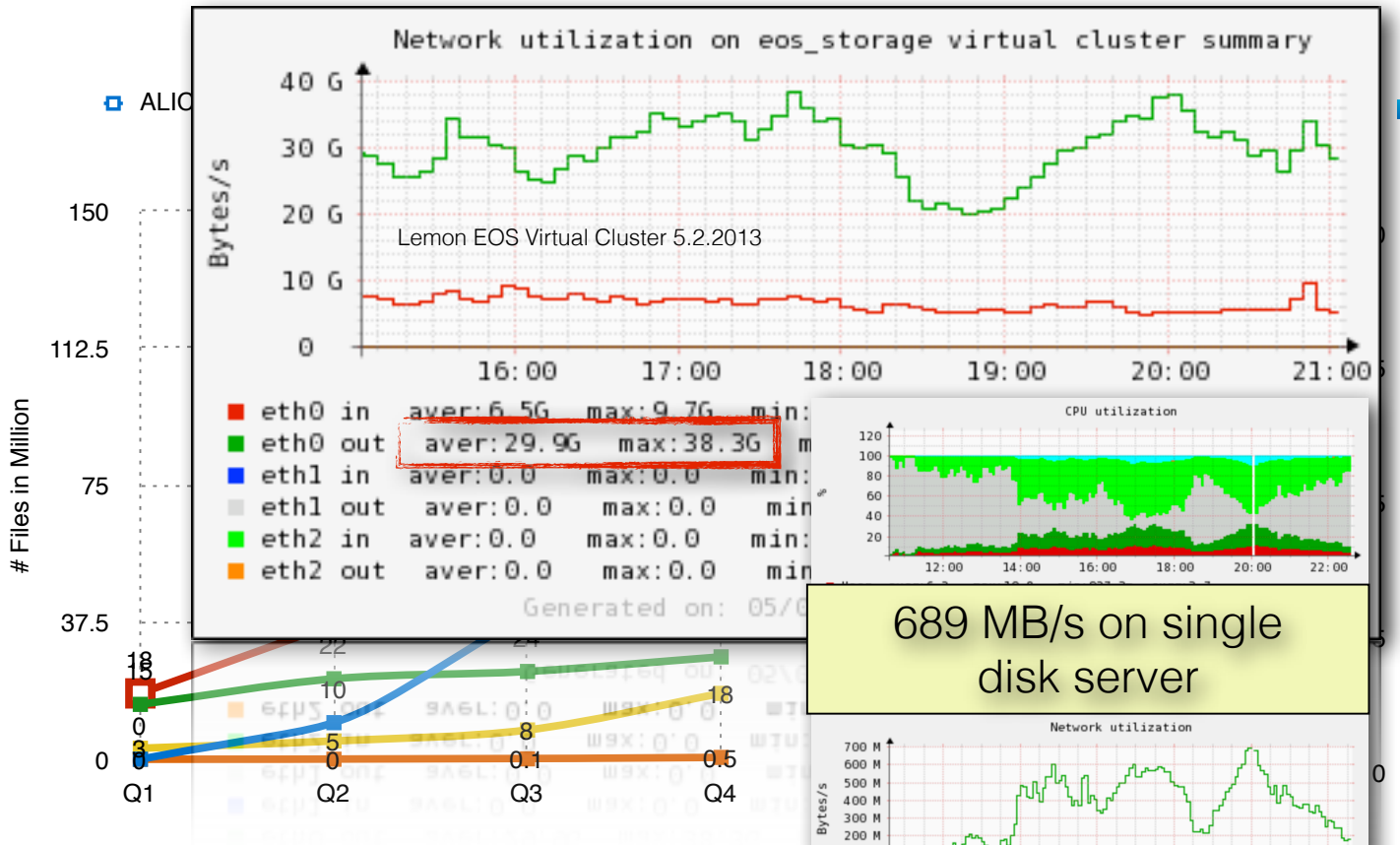


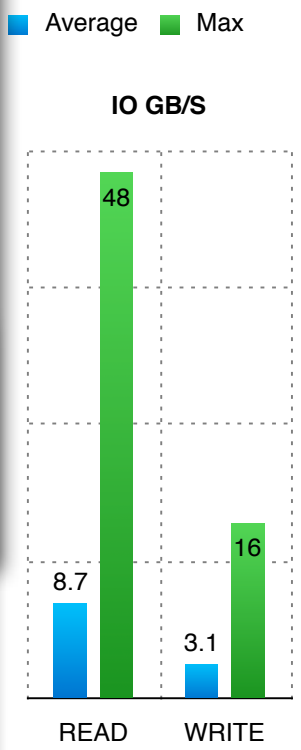
*EOSPUBLIC stated 4.6.2013

Stored Files (Replica) **136 (279) Mio.**



Service Usage





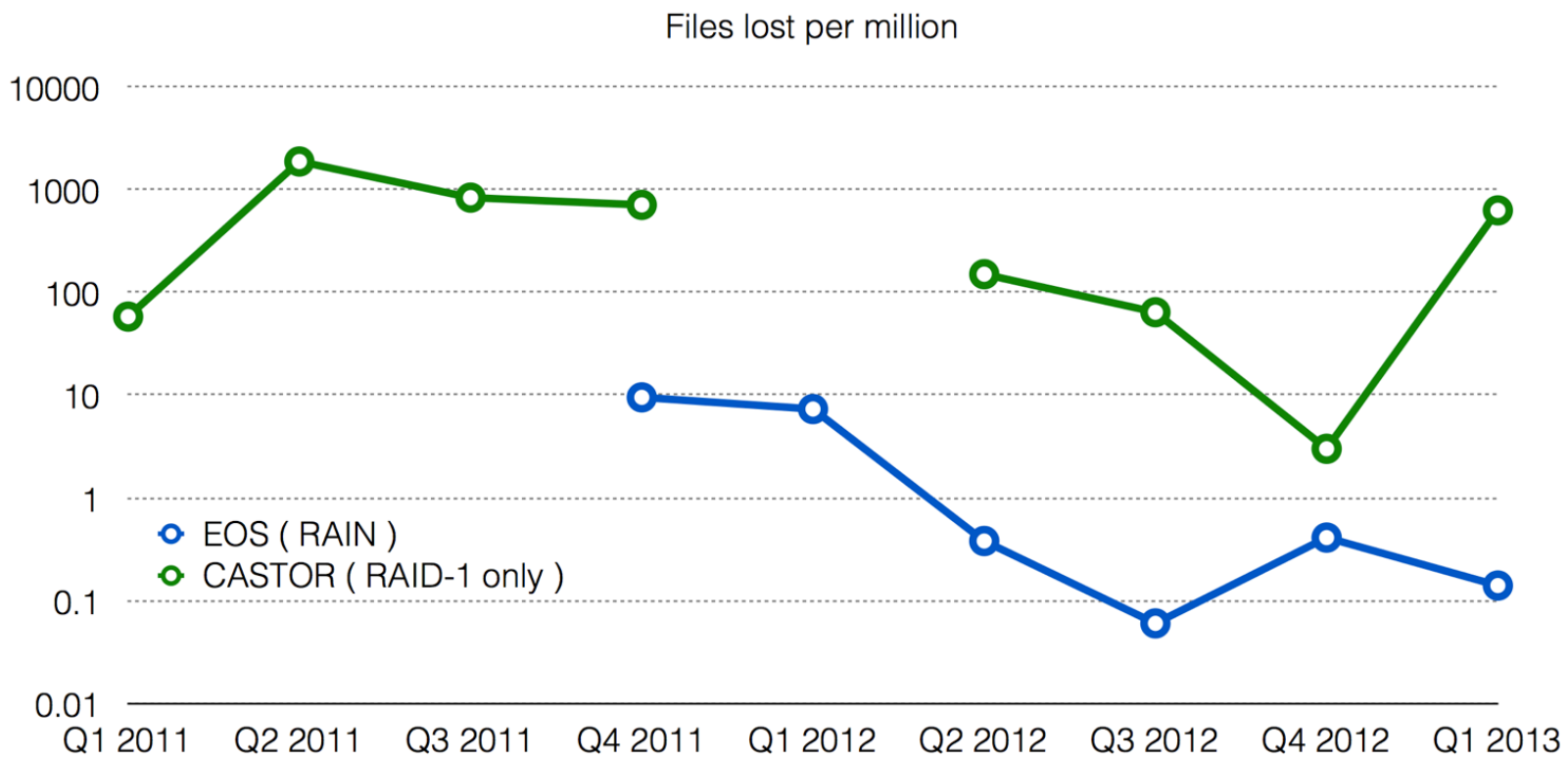
- main increase due to migration of disk-only pools from CASTOR
- IO dominated by LAN batch job access

usage

nd 26 GB/s read

- < 10% traffic volume is remote IO
- EOSALICE served 40% of ALICE IO in Dec./Jan.

File Loss Probability Comparison



File loss in EOS is mainly due to bugs and human errors - HW failures rarely lead to a file loss
File loss probability with 2 replica is not significantly higher than on tape media

Development

Software Releases

AMBER v.0.2.0 - 0.2.33

Summer 2012 - today
20 releases

BERYL v.0.3.0 - ...

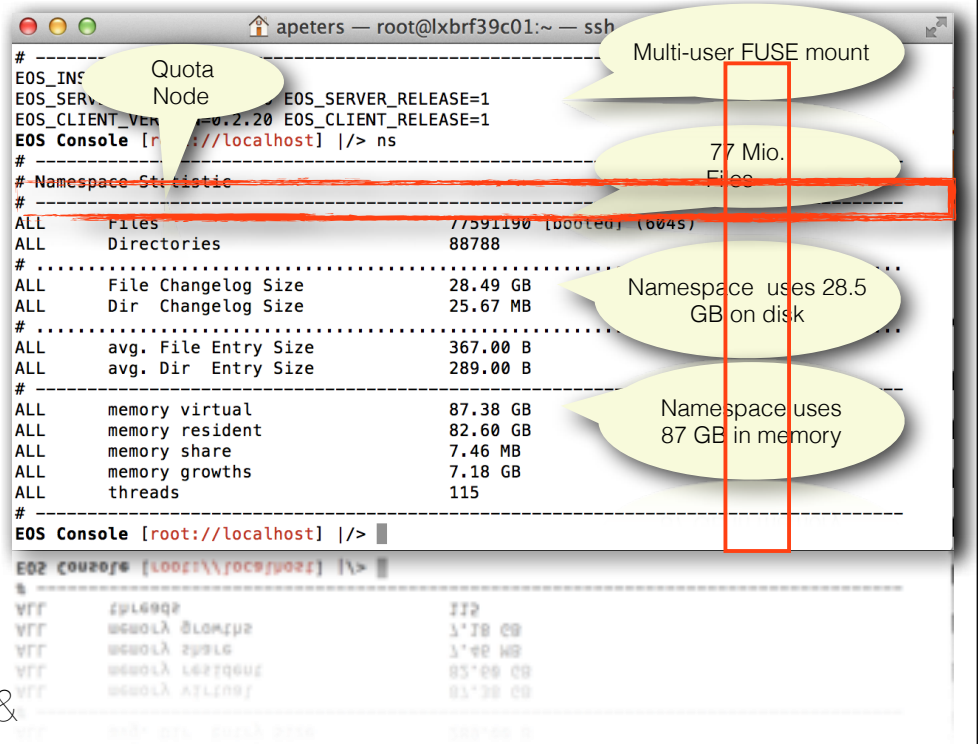
June 2013 - ...

"we are gem stones",
so plenty releases left ...

Core Features

Some core features ...

- hierarchical in-memory namespace
- POSIX-like file acces
  - XRootD protocol
  - gridFTP protocol gateway
  - FUSE single-/multi-user mount
- authentication
  - strong (Kerberos5, X509) external clients
  - shared secret (sss) internal clients
- quota system
  - user & group quota for file inodes & volume
  - quota administrator
  - quota definitions on quota nodes (namespace sub-trees)



Some core features ...

- directory user+system attributes
  - define file layout
  - define file checksum algorithm
  - define block checksum algorithm
  - define placement space
  - define minimum, maximum filesize
  - define default file pre-allocation size
  - define co-ownership by user credential
  - define ACL
- ACLs on directories
  - read, write, write-once, browse, quota, chown, chmod, deletion rights
  - defined by user, group, egroup
- automatic disk/node draining
- automatic disk balancing
- active namespace redirection & rate limiter
- real-time filesystem check
- periodic direct-IO file checksum scan

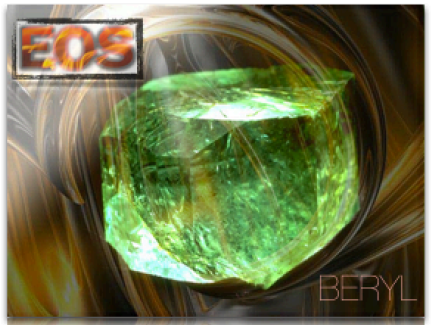
```
r  grant read permission
w  grant write permissions
x  grant browsing permission
m  grant change mode permission
!d forbid deletion of files and directories
+d overwrite a '!d' rule and allow deletion of files and directories
q  grant 'set quota' permissions on a quota node
```

```
EOS Console [root
sys.forced.blockc
sys.forced.blocks
sys.forced.checksum="adler"
sys.forced.layout="replica"
sys.forced.nstripes="2"
user.acl="u:atlas003:rw,egroup:atlas-comp-cern-storage-support:rw"
EOS Console [root://localhost] |/>
```

New in EOS BERYL

- New XRootD client
- Namespace improvements
  - Master/Slave namespaces
  - Online Compacting
- Recycle Bin
- Archive Layouts RAIN++
- Layout Conversions
- Inter-Group Balancing
- MGM node HA
- Geo Replication/Scheduling (Wigner/Cern-CC)
- 'Localhost' Scheduling (client on disk server/PROOF)
- HTTP/S3 Interface

BERYL v.0.3.0 - ...



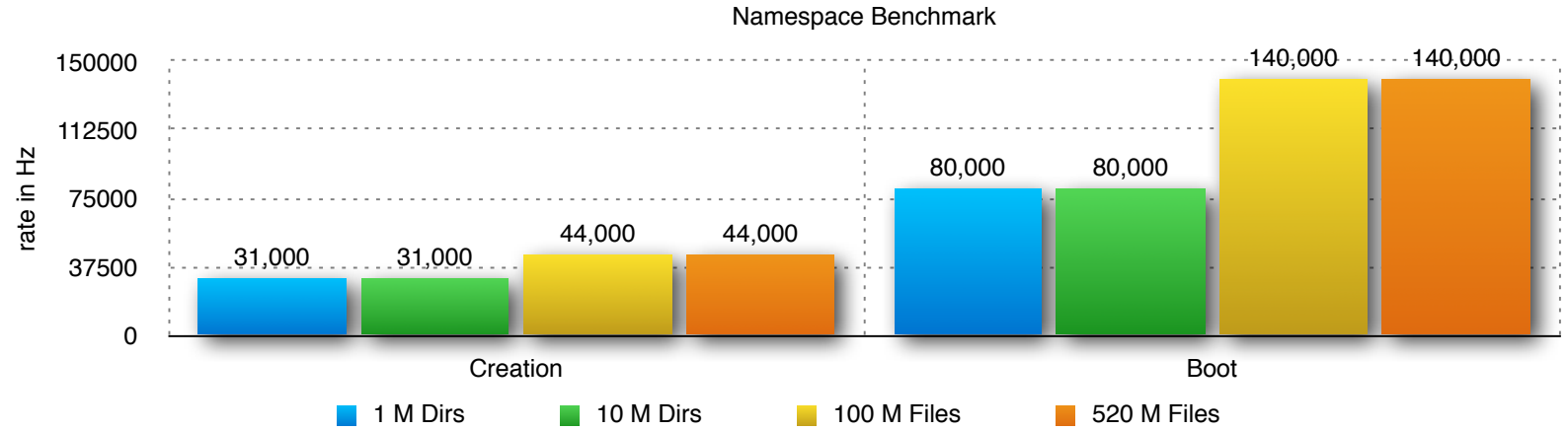
Spring 2013 - ...

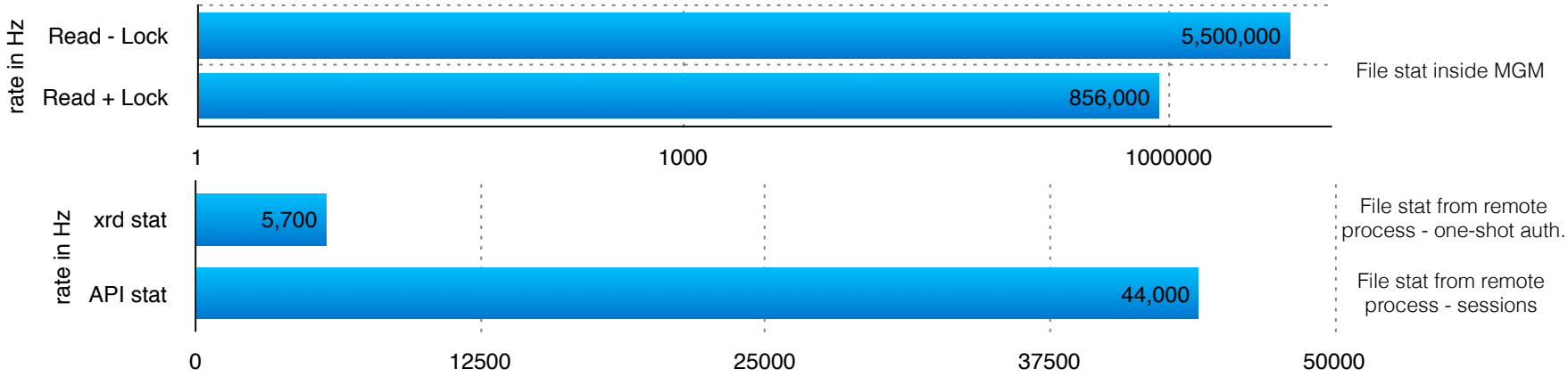
File Deletion 'Undo'

- File deletion
  - deleted files disappear immediately (seconds)
  - implemented time & volume based recycle bin

Namespace Benchmark

- MGM are equipped with 256GB memory (2x6core with Hyper Threading 2.0 GHz)



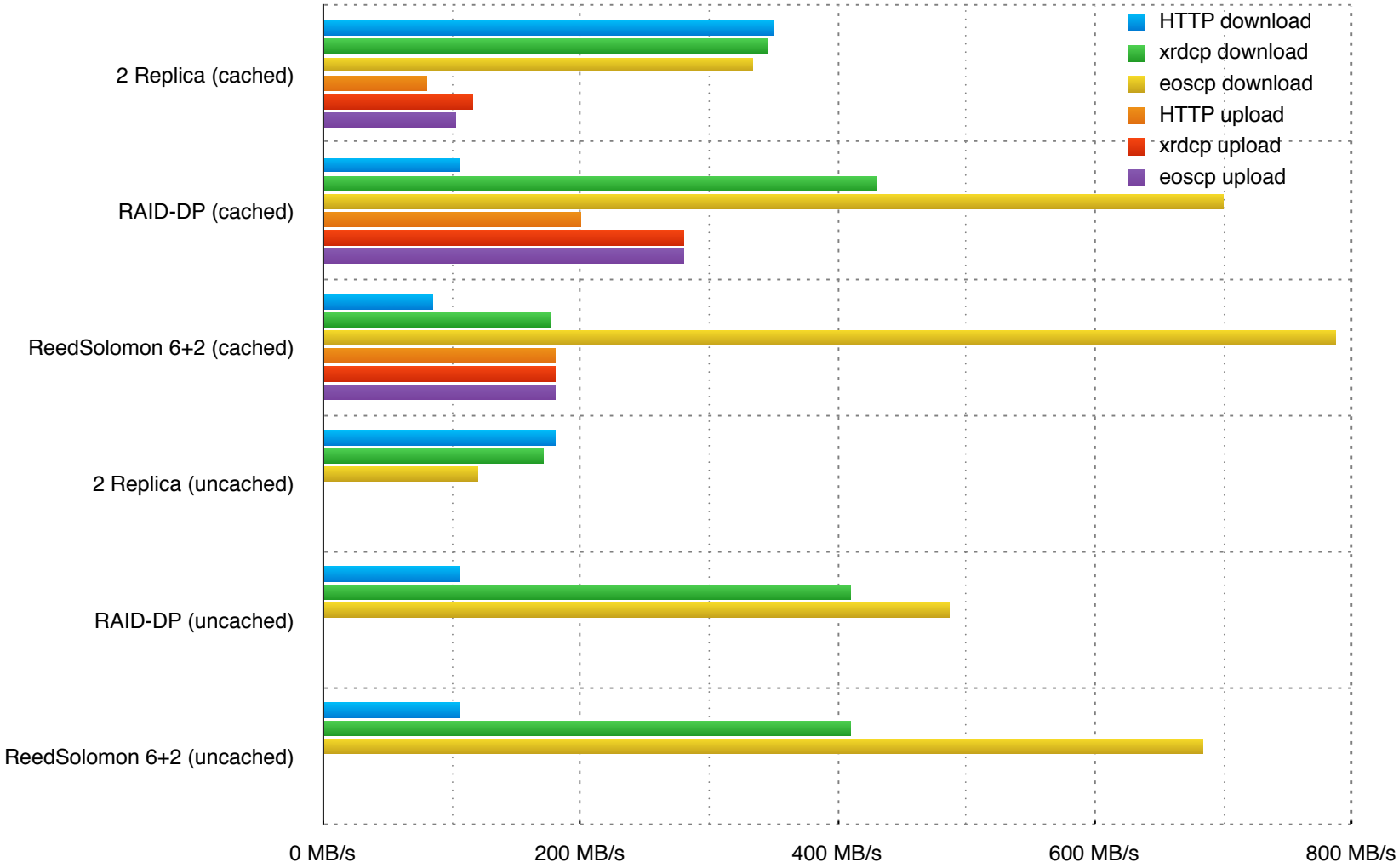


RAIN replaces RAID configurations

- Each file can have an individual striping layout
- Files are striped inside a placement group over **N+M** nodes
- EOS supports
  - Replica
    N replica of files (1 <= **N** <= 16)
  - RAID-DP
    4+2 Dual Parity Layout
  - RAID6
    N+2 RAID 6 Reed Solomon Layout (you can loose 2 disks without file loss)
  - Archive
    N+3 RAID 6 Reed Solomon Layout (you can loose 3 disks without file loss)
- Every stripe is written with a configurable blocksize and **4k** block checksums (hw accelerated **CRC32C**) which allow error detection and correction on the fly

Preliminary 10GE Benchmark

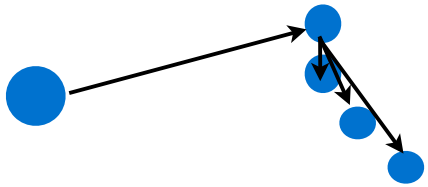
Single Client Performance against 8 standard disk server



RAIN Files for Analysis?

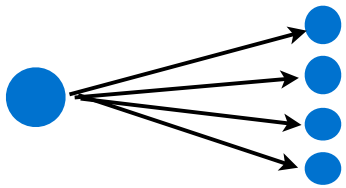
Not yet perfect ...

- As of today **RAIN** adds **additional** LAN **latency** (RT between disk server) to analysis (readV)
- Two options for high performance analysis support
  - XRootD 4.0 exposes readV call in OFS plugin
    the gateway server can read asynchronous from several remote disks boosting performance involving more disk spindles



in approx. 2 months

  - The new XrdCl will offer plugin interface with EOS IO: readV calls are asynchronously fetched from several remote server

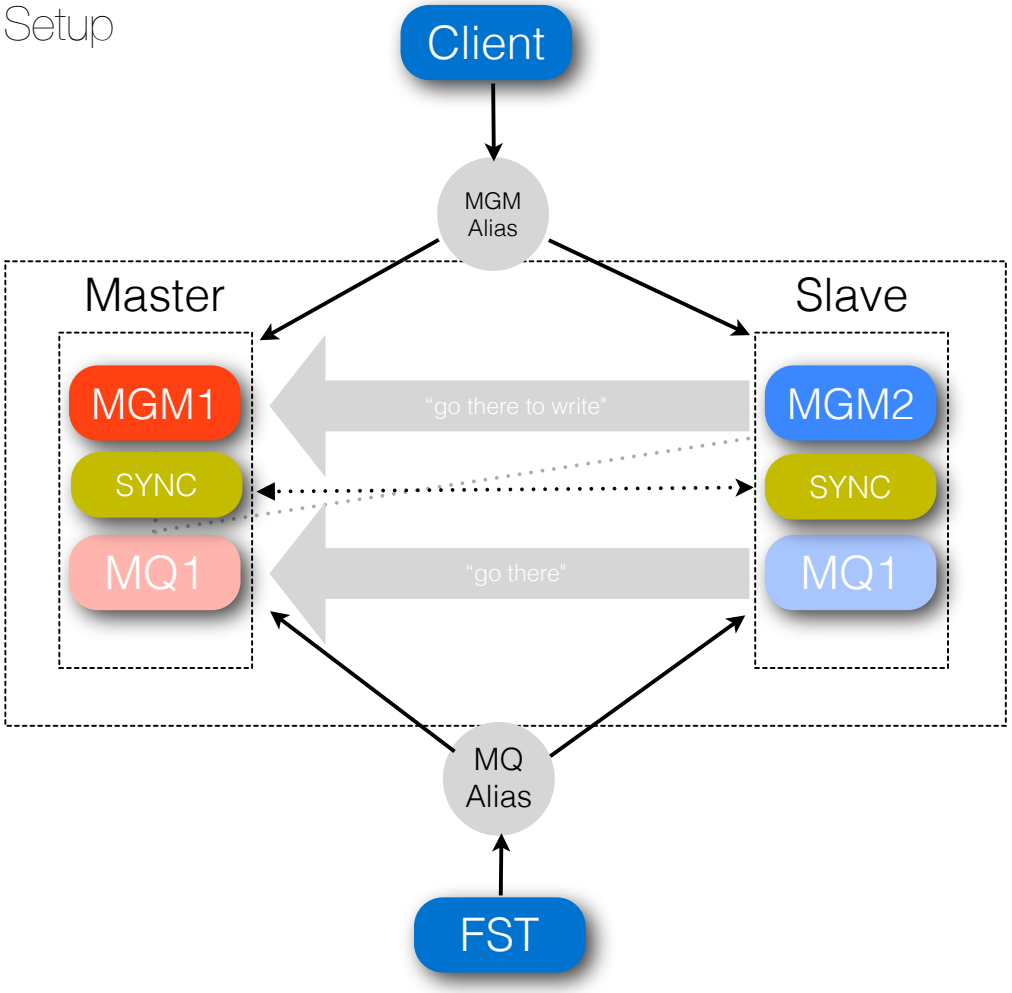


timescale undefined

HA Features

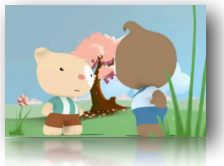
“How can we avoid/reduce downtime?”

Master/Slave Setup



Is EOS free of limitations?

Is it web scale?

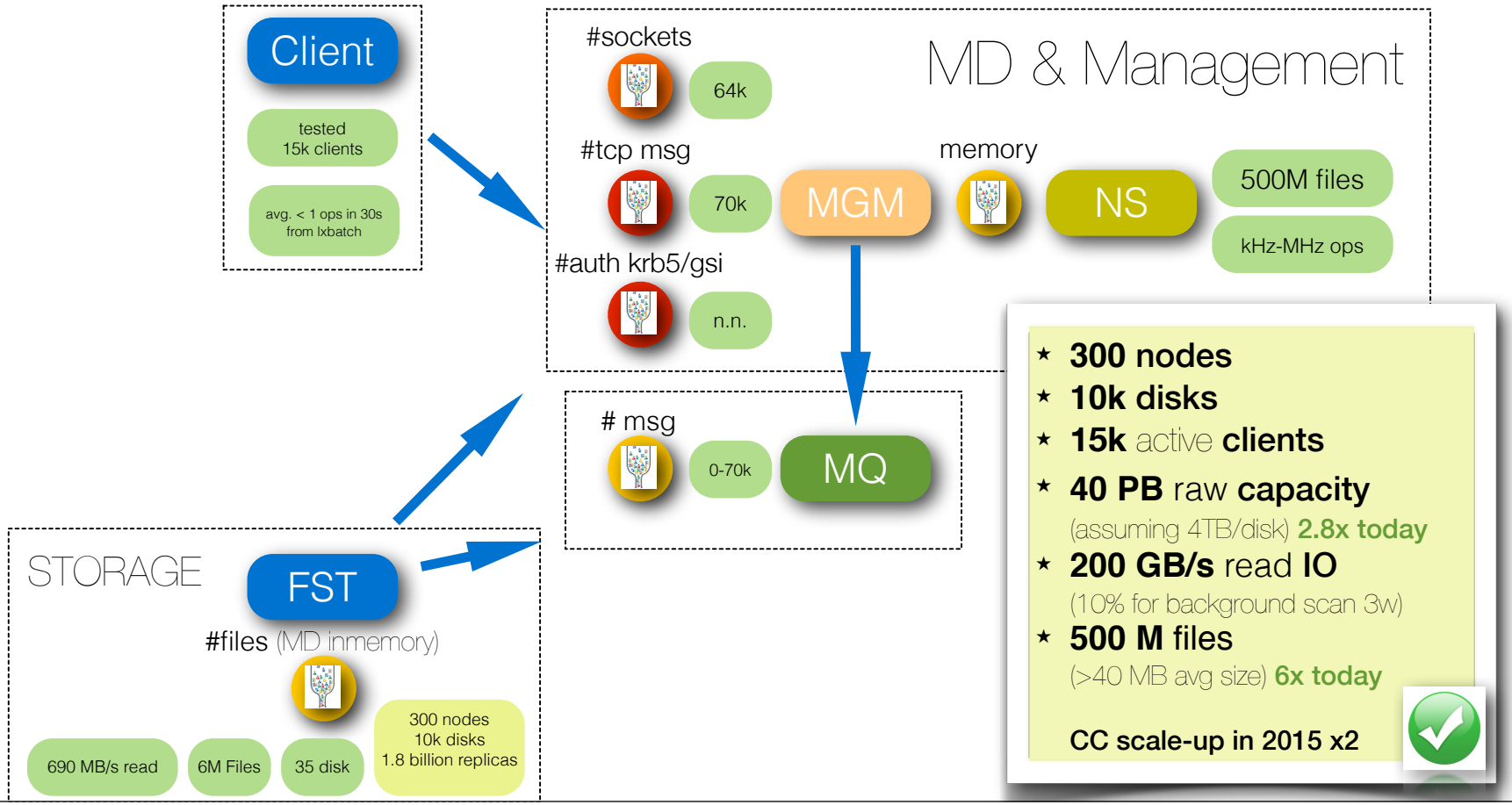


Is it really organic?

Scalability

“How big can one instance be based on what we have today?”

The performance indicators …



BERYL Release Plan

- next week target **BERYL** release **0.3.1**
  - new XrdCl, full RAIN support, Namespace HA, Recycle Bin
- following weeks of June/July
  - agile fixes + refactoring **0.3.X ...**
  - enable Layout Conversion interface & ext. balancing
  - enable HTTP(S)/WebDAV/partial S3 interface

EOS as ALICE T1/2 Storage

Compare EOS to an XRootD native installation ...

- **PRO's**
  - EOS is optimized to deal with **unreliable storage hardware** (RAIN)
  - EOS has **management functionality** for very large (inhomogeneous) installations and lifecycle management
  - EOS **tracks** storage **contents**, reports unavailable or missing files and can repair errors
  - EOS is a **multi-user storage platform**
  - EOS supports XRootD 3.3 **TPC** (third party copy mechanism)
  - EOS allows **remote administration** via EOS shell

- **CON's**
  - If you don't need any of the PRO's, it simpler to run/install XRootD => XRootD has no **stateful meta-data server** (redirector)
  - EOS can **not** be combined with **FRM** configurations
  - EOS provides **more functionality than** the **baseline** needed by an ALICE SE
  - EOS does **not support** the 'old' **xrd3cp** as a source SE

EOS Support Model

- EOS is an **OpenSource** project - usage is free at own risk
- CERN offers **best-effort** support for EOS@CERN
  - we give support to FNAL in an informal way
- EOS **does not differ** from in the support model
  - we have ~5 experts at CERN able to help with operational questions

How to move to EOS ...

- You can **move** from a **native XRootD SE to EOS** using the **same hardware**
  - use EOS AMBER release with single replica layouts on RAID arrays
  - allows to serve part or all available space to ALICE GRID with alice authentication enforced
  - additional option to have local user space with user quota with krb5/x509 authentication
  - should have 1-2 nodes with decent memory as namespace machines (1 GB RAM = 2 Mio. files)
- **Migration**
  - deploy EOS on top of existing XRootD infrastructure
    EOS uses port 1094 on MGM and default's to 1095 on storage server
  - set ENOENT redirection from EOS to XRootD and change URL in ALICE SE configuration to point to EOS
  - migrate the XRootD SE to EOS using EOS transfer queue system or xrdcp scripts
    - if a file is migrated it is served by EOS
    - if a file is not migrated EOS redirects to the XRootD redirector
    - new files are written to EOS
    - good opportunity to clean-up the SE (get a central DB dump)
  - CERN did live-migration of CASTOR/ALICEDISK to EOSALICE without service downtime

How to move to EOS ...

- You can **move** from a **native XRootD SE to EOS** using **different hardware**
  - today: use EOS AMBER release with two replica layouts on top of JBOD disk server
    100% space overhead
  - from mid June: use EOS BERYL with RAIN layouts on top of JBOD disk server
    <50% space overhead (configurable)
    - it is no problem to upgrade from AMBER to BERYL and to convert files from 2 replica to RAIN layouts
  - Migration procedure identical
  - Standard XRootD monitoring works on EOS
  - EOS ApMon RPM available

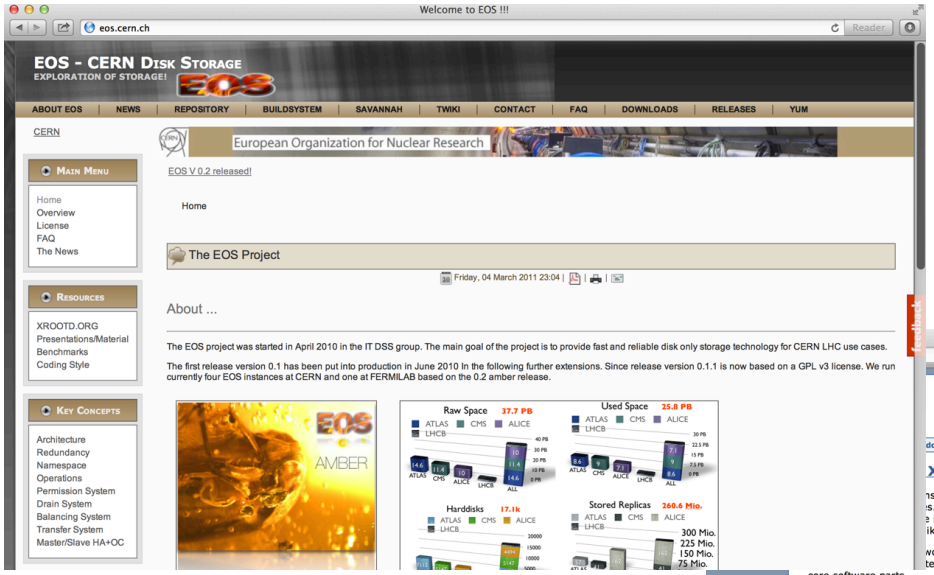
- You can easily deploy a new SE with **EOS** on empty server

How do you install EOS ...

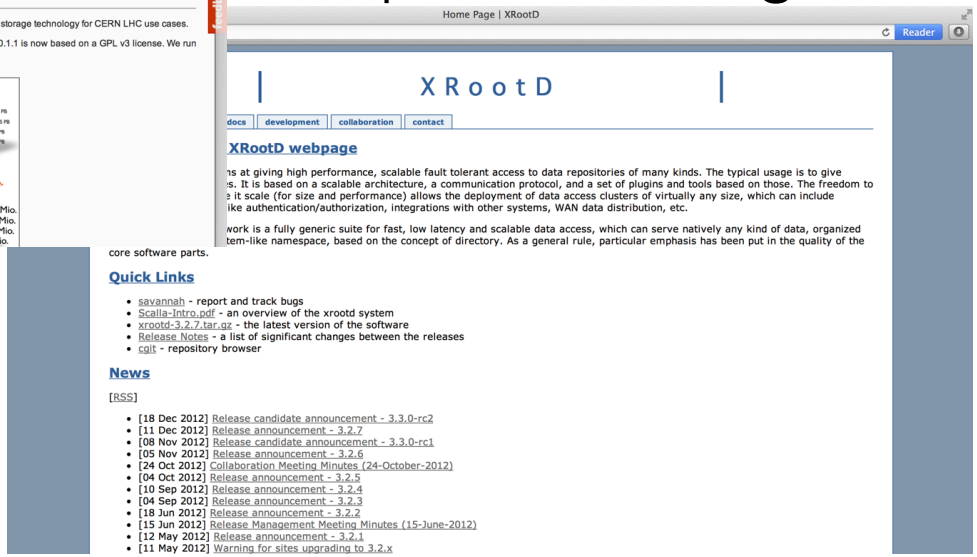
- For SLC5, SLC6 or Fedora you can use existing repositories
- There is **no** packaging (service script support) available for **Ubuntu**
  - one can convert RPMs with cpio or compile EOS from sources; still the service scripts are RedHat specific
  - contributions welcome
- You can compile EOS from sources
- Find the **HowTo** for a basic ALICE EOS-SE

Useful Information

http://eos.cern.ch



http://xrootd.org



Conclusion

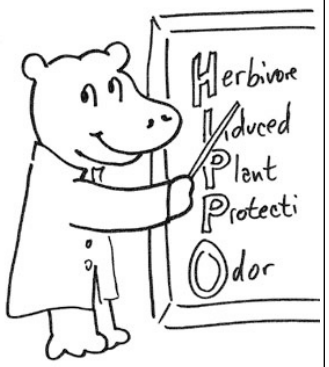
EOS Hot Storage for Cool Heads

- important storage development project in DSS
  - has become largest disk storage system at CERN during the last year
- available as Tier-1/2 (disk) storage system with full XRootD compatibility
  - a site has to evaluate if it fits its needs
  - installation,operation and configuration is simple
  - room for contributions
  - you can get in touch with EOS experts via email **eos-admins@cern.ch**

Thank You For Your Attention!

Appendix

What does EOS stand for?



pick your favorite one here ...

http://www.acronymattic.com/EOS.html

Element of Surprise
Economy of Scale
Emotion of Storage
Expiration of Service
End of Saturation
....