Andreas.Joachim.Peters@cern.ch
アンドレアス ヨアヒム ・ ピーターズ

EOS

Disk Storage @ CERN

Motivation for EOS

- separate **hot-** and **cold** storage
  - hot storage (disk) to serve user/group and experiment analysis data
  - cold storage (tape) for not-derived (raw) data
- storage for analysis use-cases
  - requires guaranteed low-latency file access
  - requires collaborative storage platform with space and access regulations
- a storage system with dynamic life-cycle management
  - expand/exchange/shrink during operation
  - semi-automatic disk ejection & healing
- simplistic deployment and operation model
  - cheap hardware & easy to use software

DSS

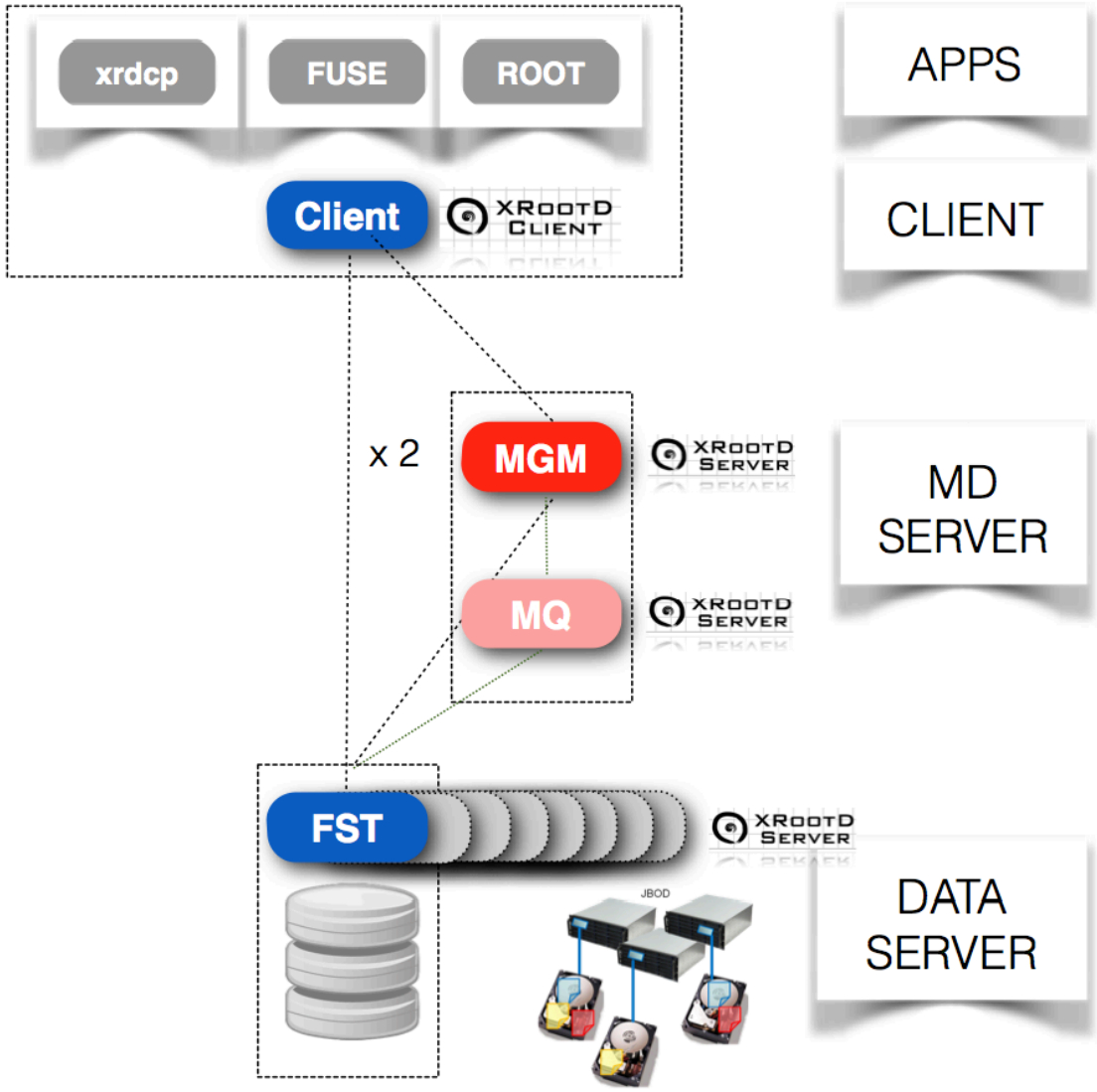
CERN Use Case

CERN IT Department

CERN disk-only file storage for (non-)LHC derived data targeting physics analysis use cases:

- five **multi-PB installations**
  (few thousand disks per instance)
  - simplified life-cycle management workflows for on-going replacement/repair of hardware
  - **JBOD** disks (no RAID controller) using software RAIN
- **low-latency** file access with in-memory namespace (ms) - 174M files
- **multi-user** platform
  - T2/T3 user/group areas with **fine-grained access control** and **quota management** e.g. shared space for the Higgs analysis group etc.
  - GRID storage element
- **secure LAN & WAN access**
  - accessed from thousands of CERN-local and remote batch nodes with Kerberos and GSI authentication
  - accessible via **XRootD**, **GridFTP**, **HTTP(S)** & **WebDav** protocol - and as a FUSE filesystem

EOS Architecture

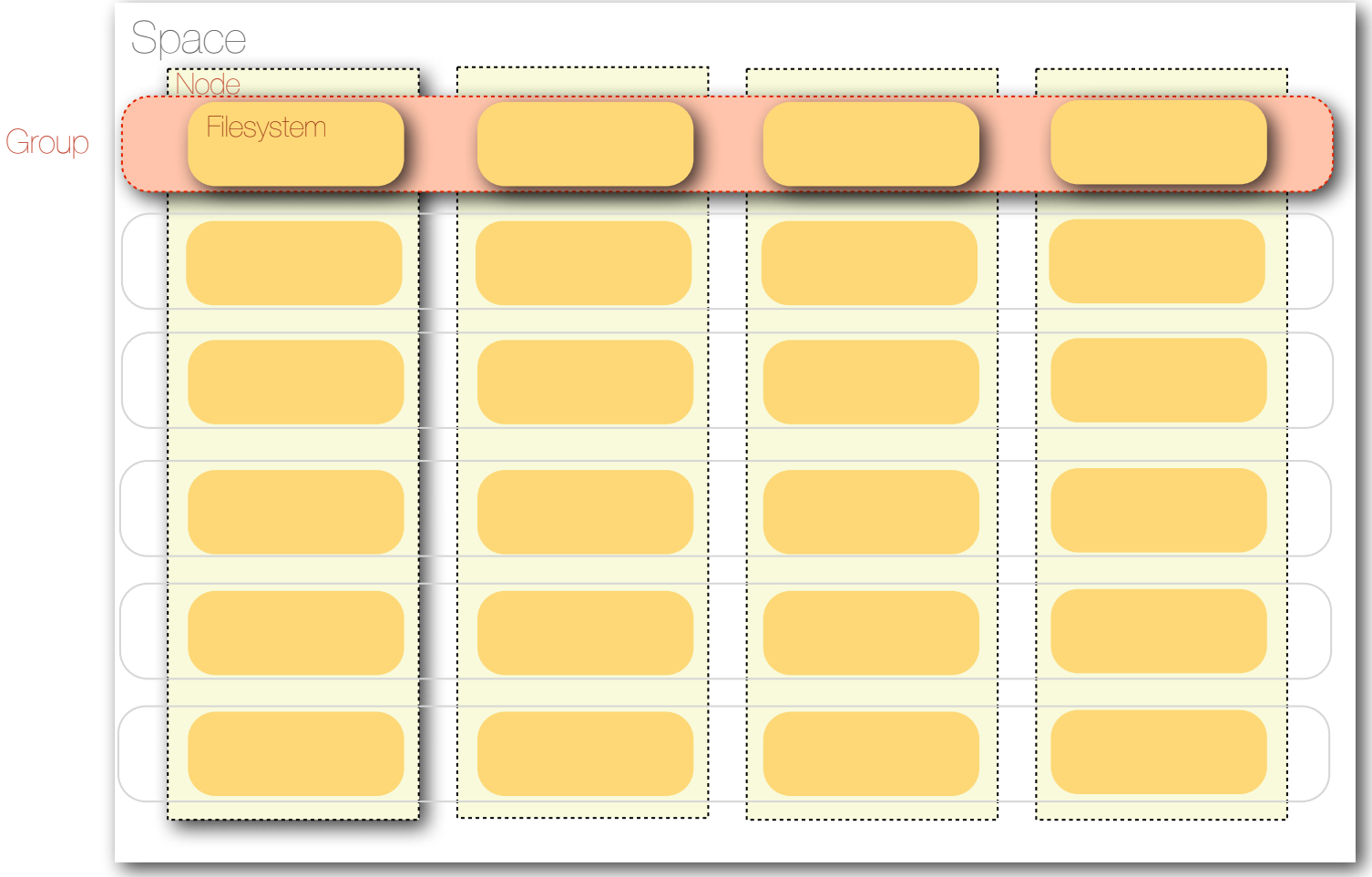


Conceptual Configuration

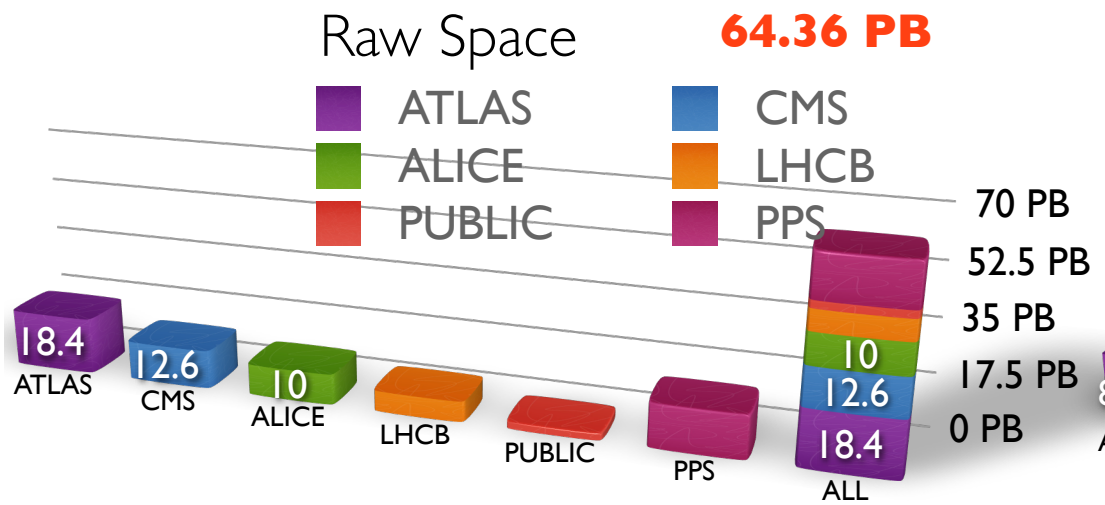
Space,Node,Group & Filesystem View

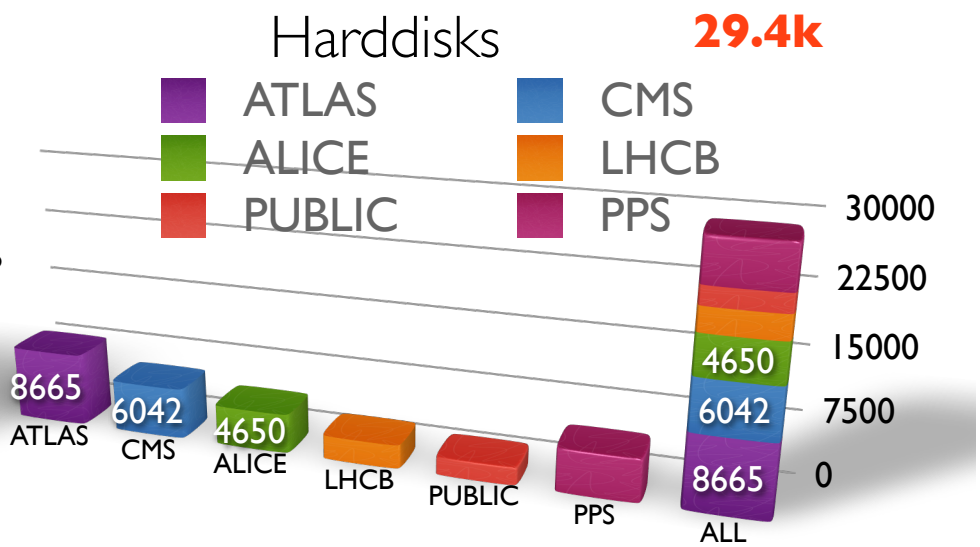
CERN IT Department

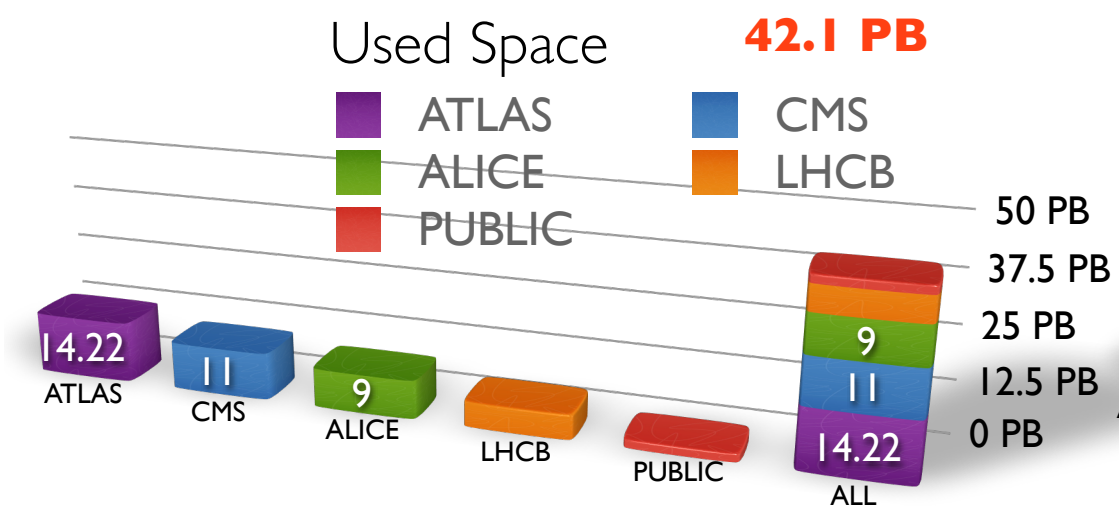
EOS organizes filesystems in views by spaces, nodes, groups and filesystems. By design there can be an arbitrary number of spaces. There should be at least as many groups as filesystems per node.

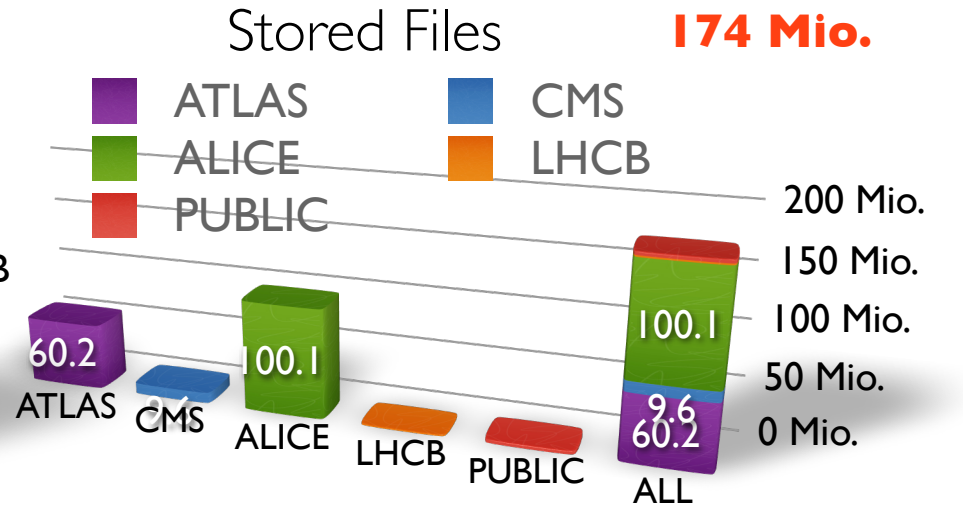


EOS Deployment at CERN









*EOSPUBLIC started 4.6.2013

EOS Deployment outside CERN

CERN IT Department

1st external instance of EOS at **FNAL** for **CMS** as Lustre/Bluearc-NFS replacement

- mainly accessed via FUSE
- single replica (RAID disks)
- 1.6 PB

2nd external instance of EOS at **SASKE**

(Martin Vala)

- 16 TB
- CAF output into volume based scratch directory
- Nanofluid/Cosmics department starting usage
- FUSE mounted on SASKE lxplus
- Evaluating PROOF on top of EOS

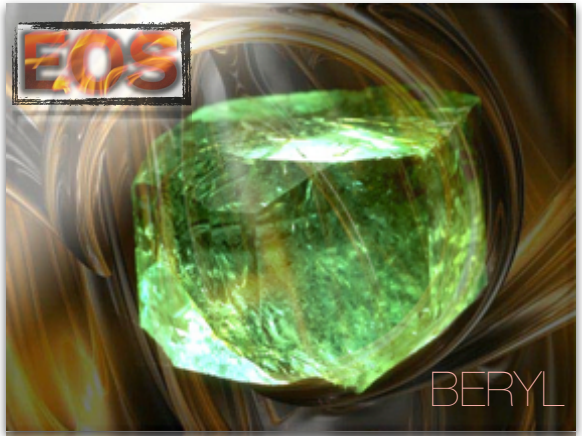
Other instances in Taiwan

(2 instances: AMS 0.4 PB, HPC as Lustre replacement 0.6PB),

Kurchatov T1

EOS Production Version

BERYL v.0.3.0 - 0.3.21



Summer 2013 - today

Addressed by EOS Beryl

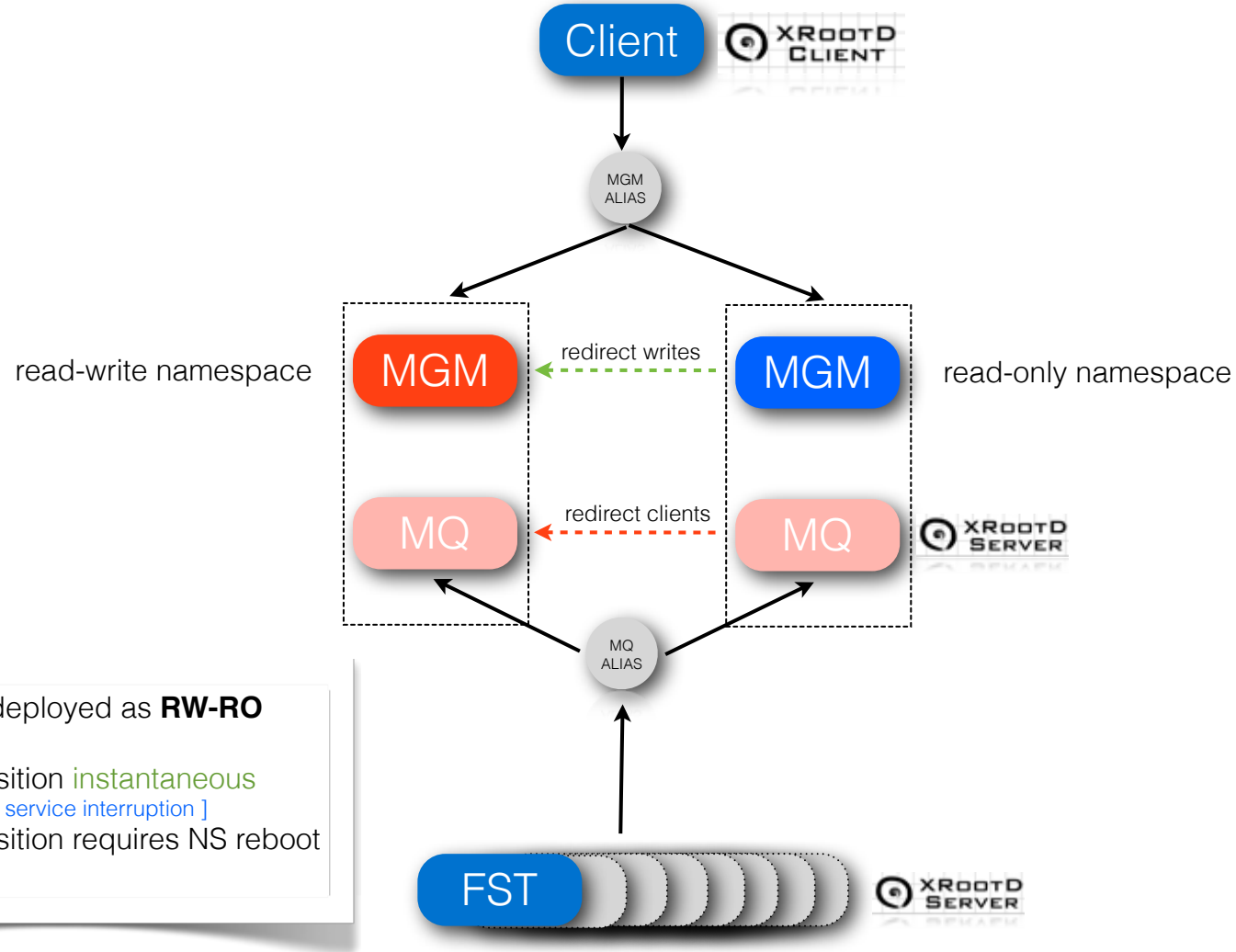
- **cover accidental deletions**
  recycle bin
- **improve reliability/high availability**
  Master/Slave namespace
- **decrease cost, increase reliability**
  ECC erasure encoding/RAIN e.g. (4,2) : 50% space overhead
  LRU cache & policy based file conversion e.g. after 1 month from 2 replica to (8,2) : 25 % space overhead
- **client for multithreaded applications**
  new XRootD client
- **integrate remote CC according to IT planning**
  GEO replication support Wigner/CERN
- **add standard interface**
  WebDAV/HTTPS support with KRB5 + X509 authentication

Addressed by EOS Beryl

- **provide/improve POSIX-like client**
  FUSE reimplementation/multi-threading/stability improvements
- **provide XRootD interoperability**
  third party copy support - today can copy @CERN: EOS <=> CASTOR
- **project quota**
  shared quota in a namespace subtree e.g. restrict the ALICE quota ;-)
- **provide generic usable GSI interface for XRootD**
  gridFTP DSI plugin for XRootD - not for ALICE!

EOS Beryl Deployment Model

High Availability Namespace Setup



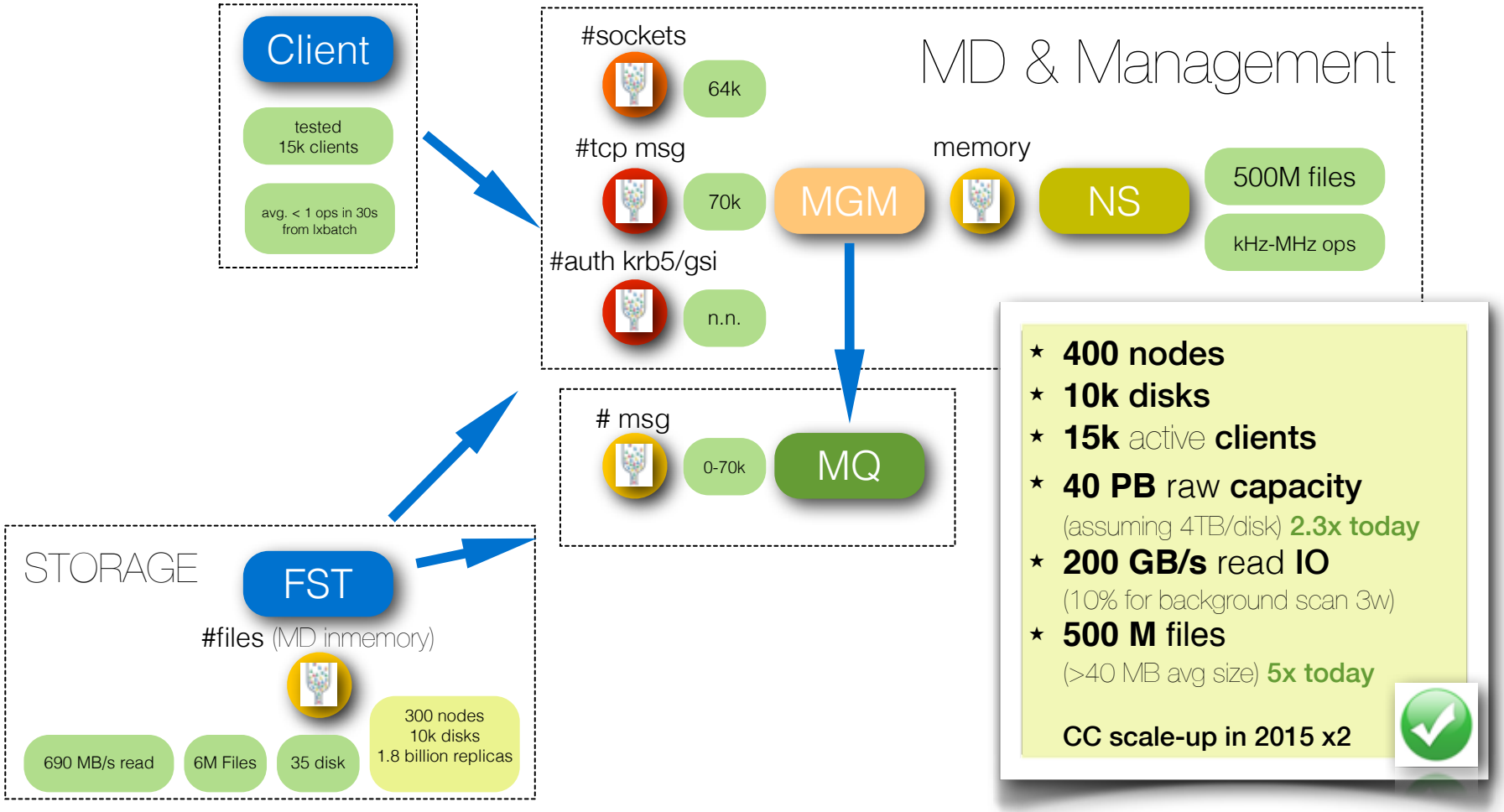
- Namespace deployed as **RW-RO** pair
- RO->RW transition instantaneous
  [HA failover without service interruption]
- RW->RO transition requires NS reboot

100-350 nodes / 2k-8k disks

EOS Beryl Scalability

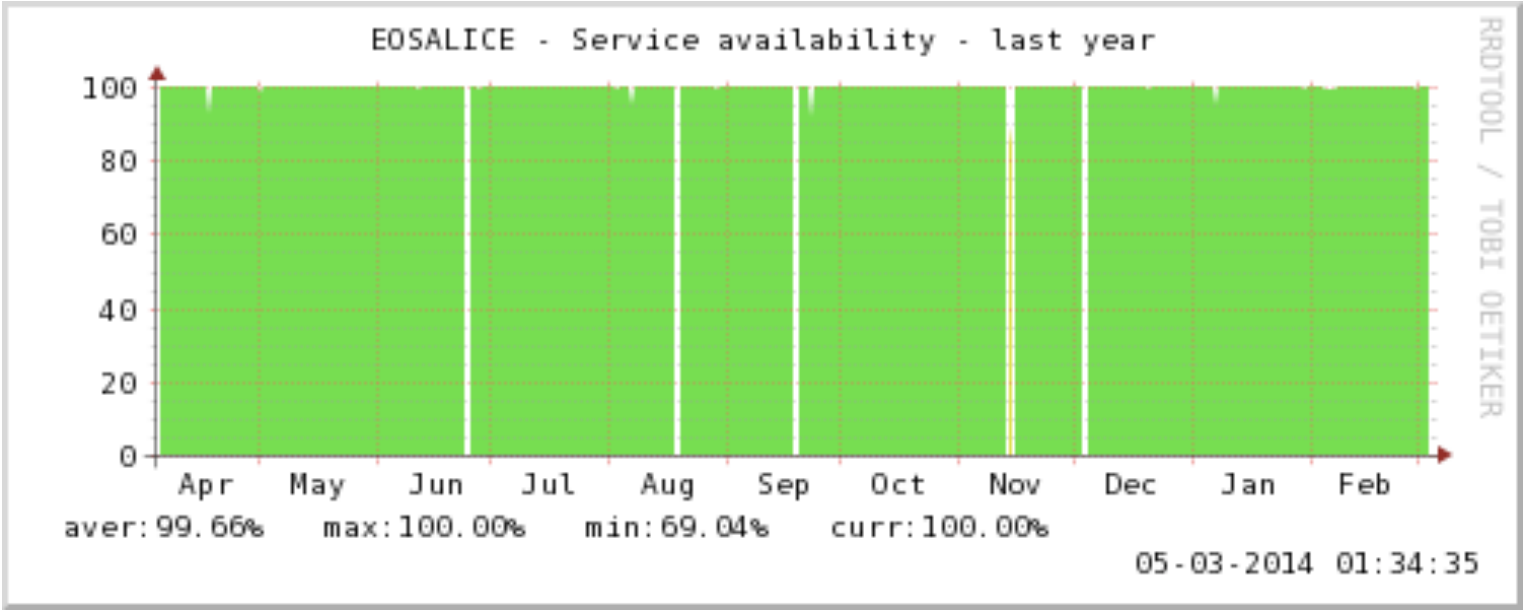
Expected 2015 capacity in largest instance

The performance indicators …



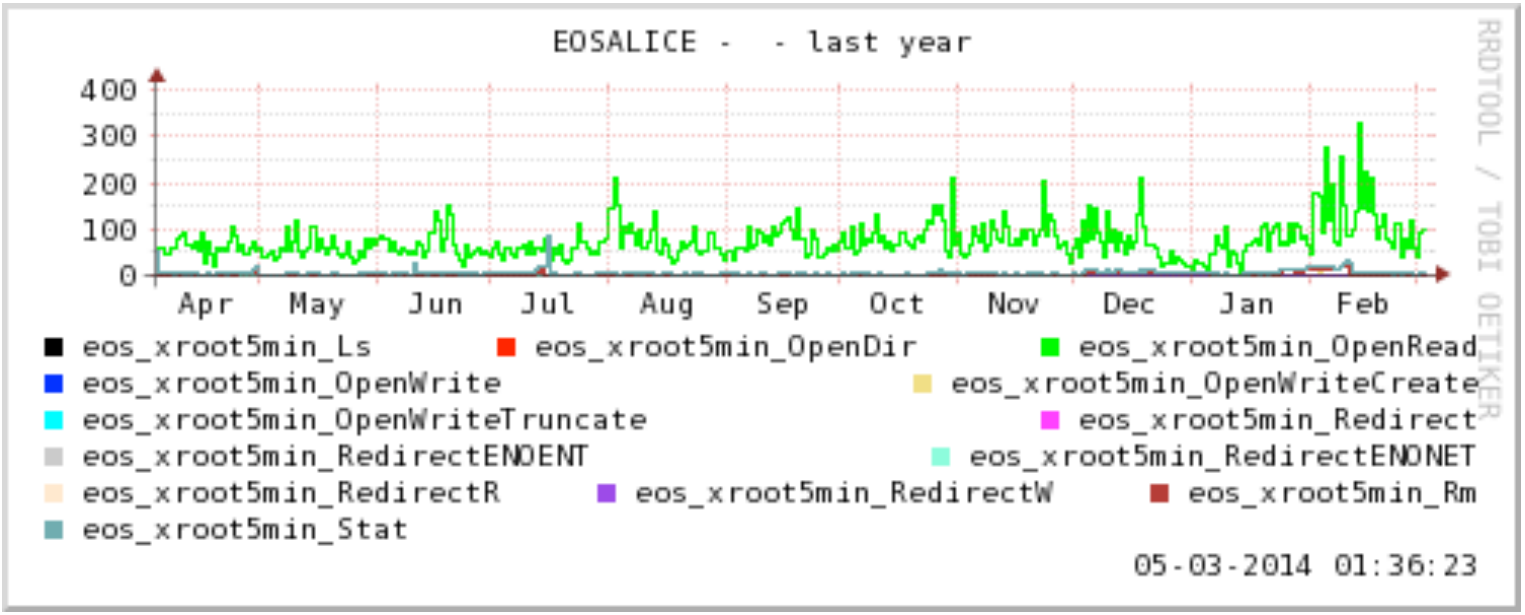
EOS ALICE Availability & Usage

one year view including scheduled downtimes



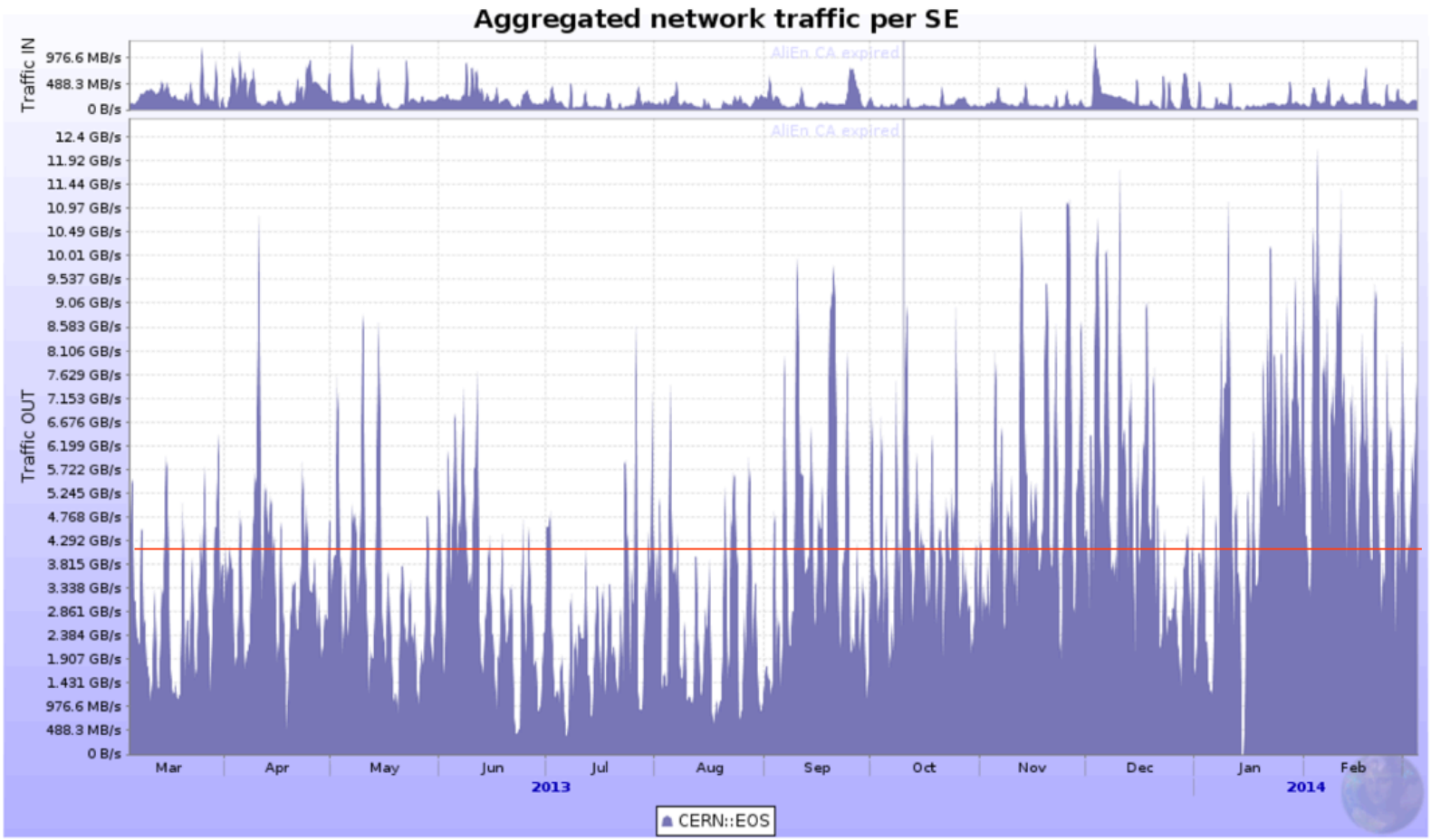
99.66% availability

- two updates
- several times reverse DNS errors originating from indian site
- two headnode software crashes
- 30 hours total downtime/ year



EOS ALICE Usage

one year view



EOS

Traffic IN

	Series	Last value	Min	Avg	Max	Total
1.	CERN::EOS	175.1 MB/s	6.587 B/s	202.4 MB/s	7.057 GB/s	5.937 PB
	Total	**175.1 MB/s**		**202.4 MB/s**		**5.937 PB**

Traffic OUT

	Series	Last value	Min	Avg	Max	Total
1.	CERN::EOS	7.511 GB/s	0 B/s	4.125 GB/s	31.98 GB/s	123.9 PB
	Total	**7.511 GB/s**		**4.125 GB/s**		**123.9 PB**

ALL

Total	**1.17 GB/s**		**940.8 MB/s**		**27.37 PB**

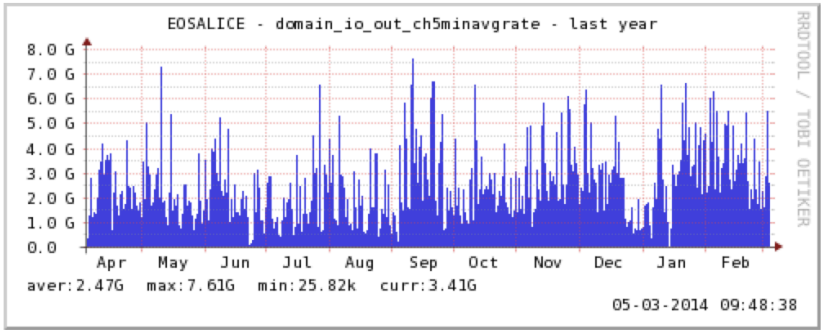
Total	**11.82 GB/s**		**8.01 GB/s**		**240.4 PB**

>50% of ALICE read IO by CERN::EOS

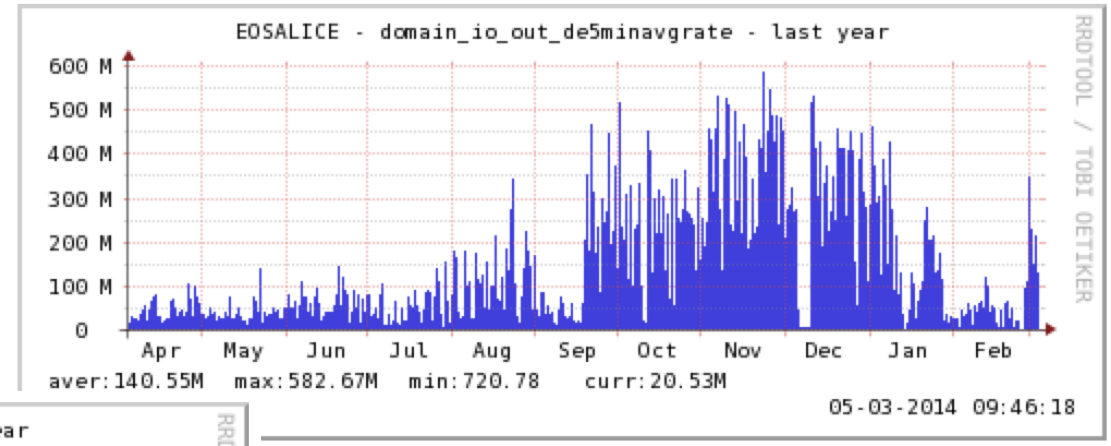
EOS ALICE Usage

Internal & External Traffic

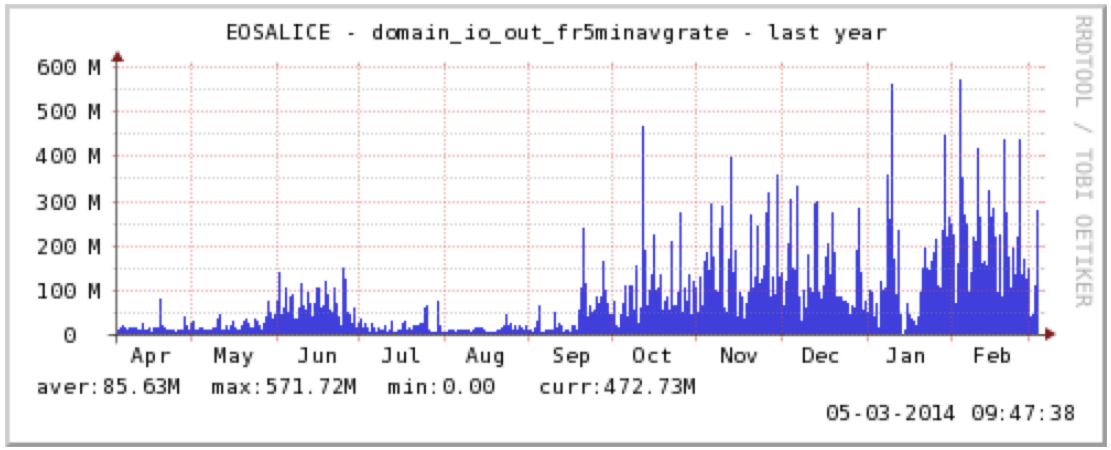

```
[root@lxbrf39c01 ~]# eos -b io stat -d
# ------------------------------------------------------------------------------------
# IO by domain/node name:
# ------------------------------------------------------------------------------------
io          domain                                    1min      5min        1h       24h
# ------------------------------------------------------------------------------------
OUT         .ro                                     1.38 G    6.91 G  187.49 G   10.55 T
OUT         .ru                                    53.71 M  223.72 M   40.94 G   16.86 T
OUT         .fr                                    14.55 G   76.06 G    1.04 T   61.85 T
OUT         .no                                       0.00      0.00      0.00  100.61 k
OUT         .cz                                     4.06 G   14.85 G  305.72 G   22.88 T
OUT         .uk                                   948.71 M    3.97 G   79.49 G  750.35 G
OUT         .su                                    63.77 M  535.22 M   23.83 G  874.20 G
OUT         eos                                     6.64 G   27.34 G  371.87 G   17.99 T
OUT         .dk                                    36.10 k  348.58 k    1.34 G   94.06 G
OUT         .org                                  689.93 M    7.48 G  170.42 G    9.69 T
OUT         lxplus                                    0.00      0.0
OUT         lxb                                    30.19 G  173.57
OUT         .nl                                     1.45 k    1.45
OUT         .it                                    20.38 G   90.76
OUT         other                                  15.03 G  115.40
OUT         pb-d-128-141                              0.00      0.0
OUT         .se                                       0.00      0.0
OUT         .ch                                    48.38 G  273.58
OUT         .edu                                      0.00  429.67
OUT         .de                                     1.35 G   11.25
```



Traffic at CERN



Traffic to Germany

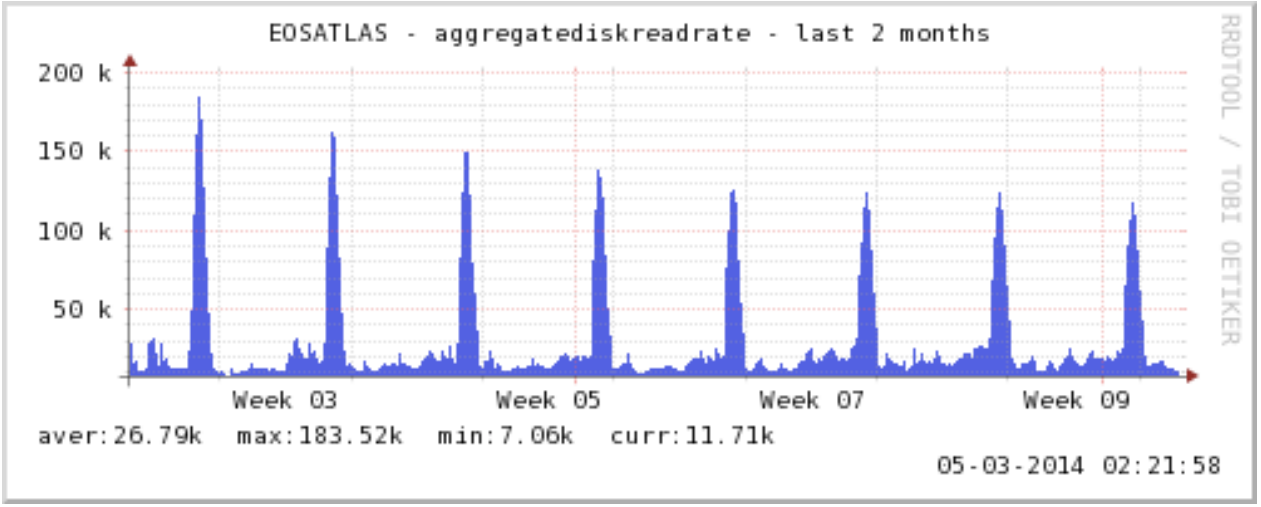


Traffic to France

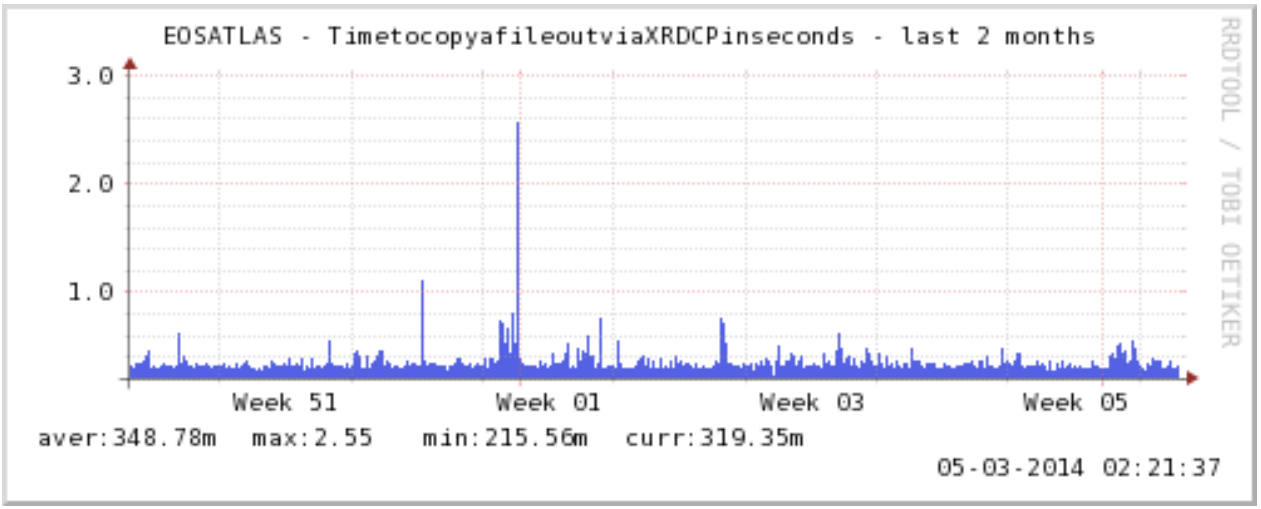
EOS Storage Content Validation

Load-aware File Scanning

Example of weekly file scanning in EOSATLAS



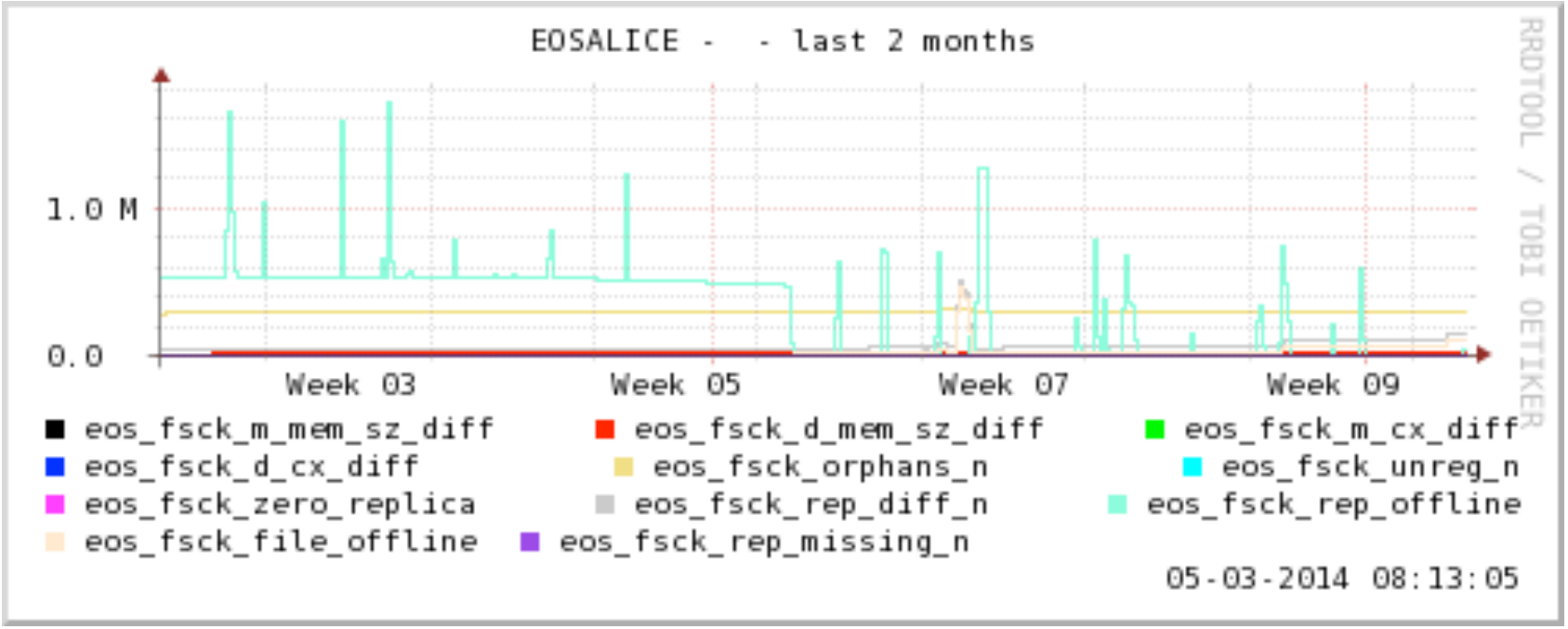
All data on EOS is read once a week using direct IO and block and file checksums are verified and errors reported. In ATLAS scanning is synchronized over all disks (in ALICE it is smeared) peaking at 100-200 GB/s.



There is no correlation(degradation) of file access latency or IO rates with the file background scan.

EOS Storage Content Validation

FSCK Results & Disk Probe



FSCK tool reports scan reports: unavailable files, inconsistencies etc.

EOS probes filesystems on ten physical locations every 5 minutes writing and reading a bit pattern. In case of errors filesystems get automatically disabled and if configured drained after a grace period.

DSS Current problems ...

CERN IT Department

We have identified **two GBLIC bugs** by now

- although it has biggest namespace - EOSALICE is one of the most stable instance
  - this is due to the authentication configuration
  - there is only UNIX (=no) authentication configured in combination with ALICE token authorization
  - we identified the major reason for instabilities in ATLAS/CMS instances in the past week
    - KRB5 and GSI Authz in XRootD uses **seteuid/gid calls**
      - dead locks if a thread is created/destroyed at the same time
      - reduces auth/s performance by factor 10!
      - this is a (known) GLIBC bug
      - XRootD 4.0 does not use anymore seteuid, there is only a local patch for XRootD 3.3.6 available

we validated that UNIX semaphores are not thread safe (bug filed to RedHat)

- XRootD 4.0 (client) avoids semaphores
- local CERN patch for XRootD 3.3.6 available

Current problems ...

Currently observing socket leaks introduced by clients running on virtualized batch nodes

- reason is not understood
- MGM/FST has established connection
- there is no connection on client side visible anymore
- work-arounds
  - configure keep-alive in XRootD
  - configure idle timeouts for connections

```
xrd.network keepalive
xrd.timeout idle 120
```

EOS Coming Releases

Future Releases/Bundle

CITRINE v.0.4.0

2014

DIAMOND Bundle

2014/15

CITRINE Roadmap

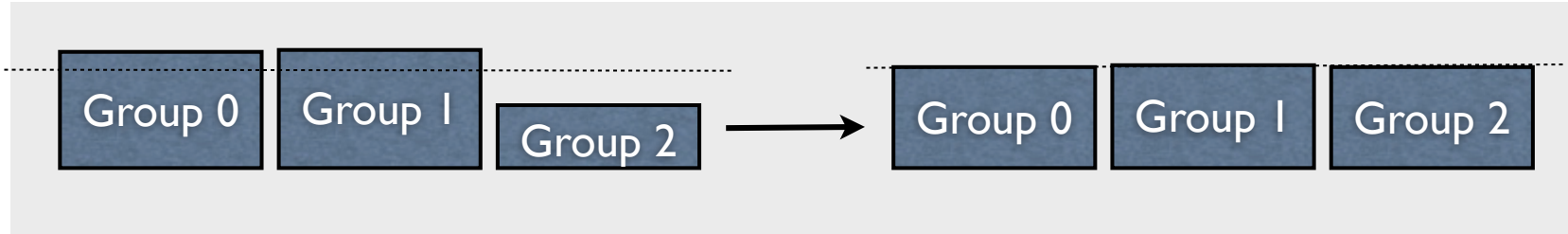
- Inter Group & Geo Balancing
- Scale-Out authentication
- XRootD 4
  - ReadV support with RAIN files
  - Thread-private authentication protocol list

    will allow to distribute a transfer sss keys to do authenticated third-party transfers from any XRootD 4.0 storage (probably not dCache)
- Topology aware Scheduling & Placement

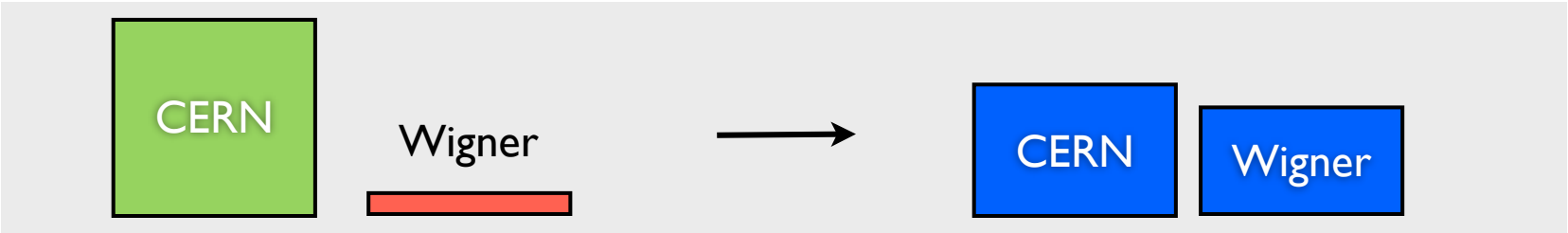
CITRINE Geo Balancing

"improve service resizing & data access efficiency in distributed CC"

- **Inter Group Balancing**
  steady pool expansion requires to add placement groups



- **Geo Rebalancing**
  optimize access in distributed CC (CERN/Wigner)



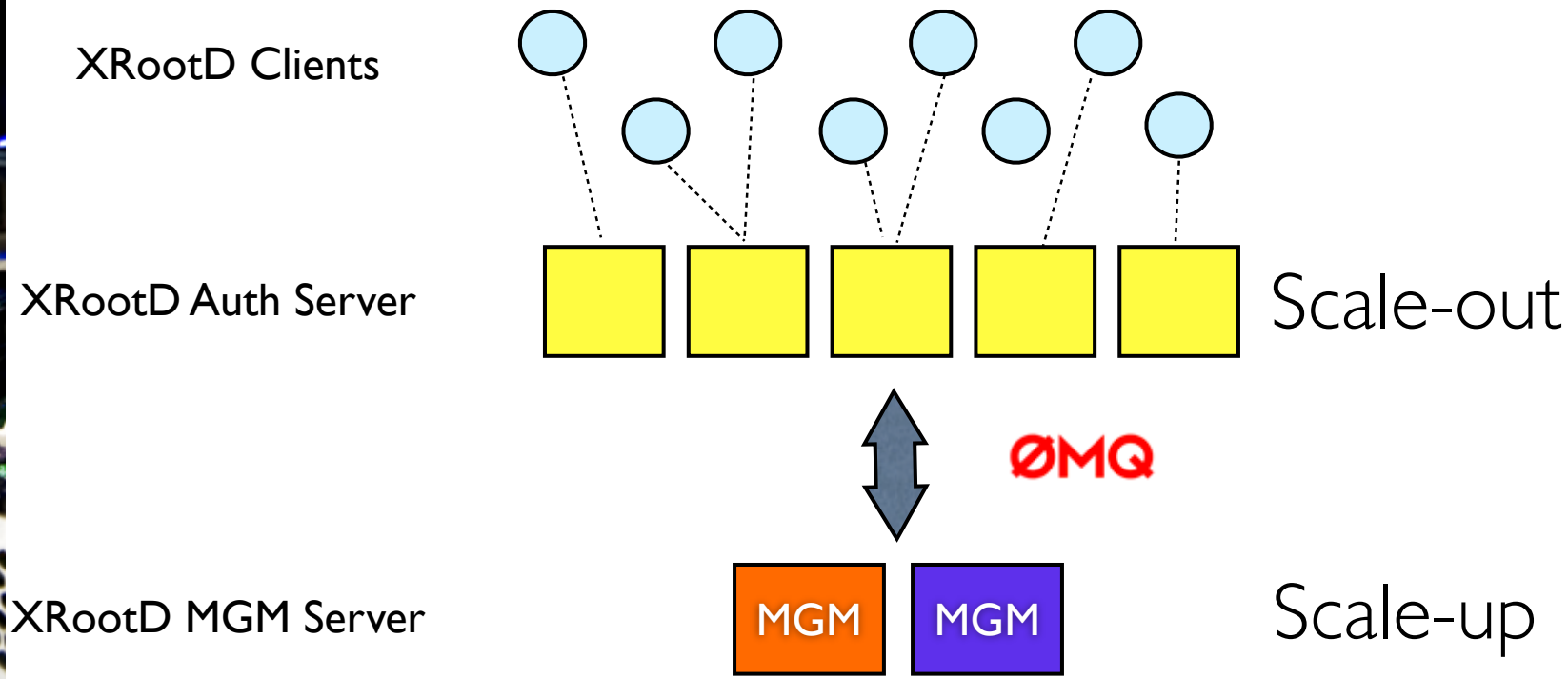
CERN IT Department
CH-1211 Geneva 23
Switzerland
www.cern.ch/it

CITRINE Scale-Out Auth

"improve scalability when increasing the number of clients (batch virtualization)"

- **Front-end Authentication**

  scale-out X509/KRB5 authentication, multiplex sockets

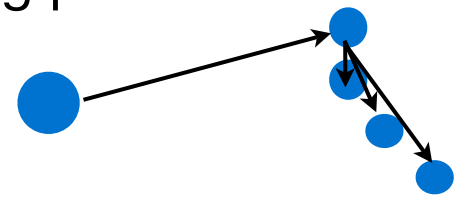


CERN IT Department
CH-1211 Geneva 23
Switzerland
www.cern.ch/it

CITRINE ReadV Support

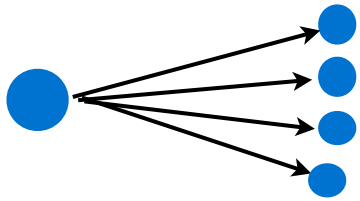
"improve analysis data access efficiency with reduced storage costs"

- 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



Done!

- The new XrdCl provides a plugin interface allowing EOS IO: readV calls are asynchronously fetched from several remote server



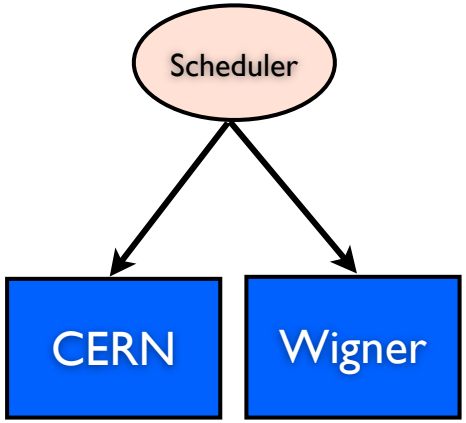
~3 month

CERN IT Department
CH-1211 Geneva 23
Switzerland
www.cern.ch/it

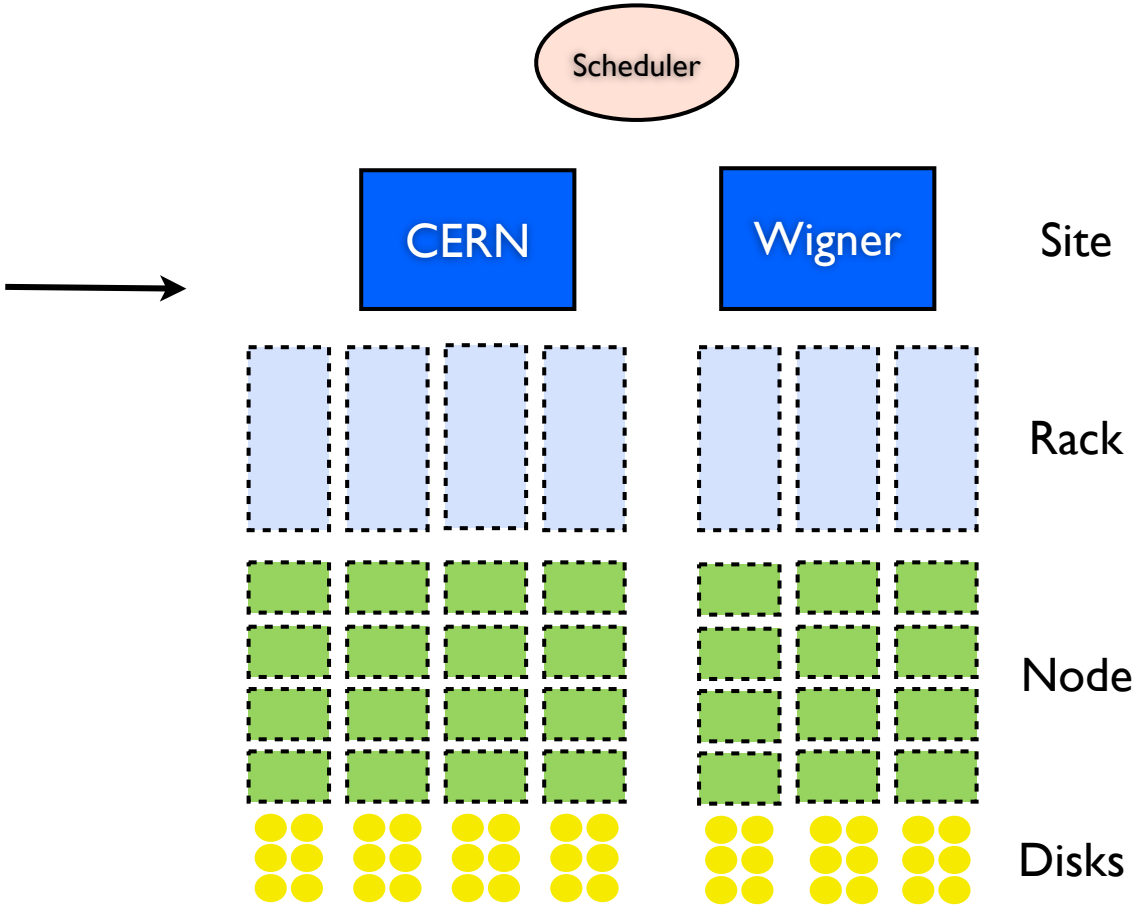
CITRINE

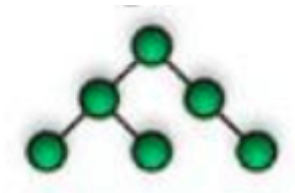
Topology-aware placement & scheduling

"reduce failure modes & improve data access efficiency"



GEO Scheduling
today BERYL





CERN IT Department
CH-1211 Geneva 23
Switzerland
www.cern.ch/it

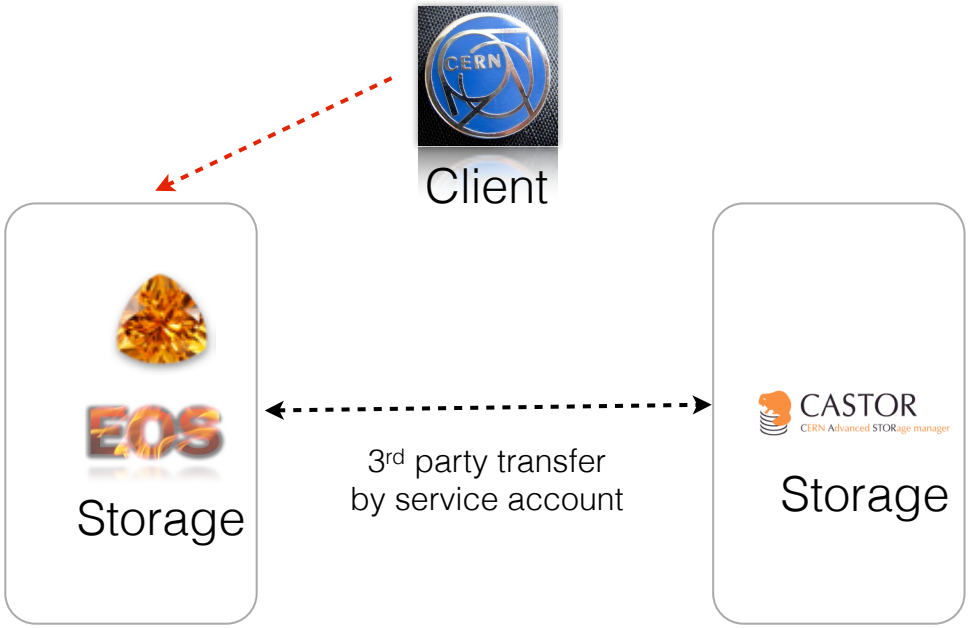
CITRINE Roadmap Schedule t.b.d.

- Beryl 0.3.21 — **today**
  - Merge new EOS shell
  - Merge ReadV + XRootD 4 (pending release)
  - Merge LevelDB FST MD
  - Log Format Streamlining
- Citrine 0.4 — **March 2014**
  - Topology Aware Scheduling
  - Refactored Draining
  - Strong Auth in FUSE
  - FUSE Evolution
- Citrine freezed/deployed — **June 2014**

CITRINE Archive Interface

Current Discussion

- Currently drafting a proposal (orig. request from CMS) to add archiving functionality from EOS to CASTOR
  - for non-GRID use cases (analysis groups, small VOs)
  - archive subtree of EOS namespace
  - no direct user access to CASTOR anymore - only via archiving service integrated into EOS shell
  - archiving of files + meta data



Looking beyond EOS
DSS R&D

DIAMOND Bundle

Diamond R&D – Why do we do that?

- things we didn't achieve (by design) with EOS or DPM
  - scale-out meta data
  - SPOF/DPOF-free
  - fully organic self-healing storage - plug & forget
  - AFS-replacing storage system anytime in the future ...
- things we didn't achieve (by design) with CASTOR
  - scale-out meta data
  - modular & decoupled disk/tape stack
  - exportable/simple community software
  - based only on free software
- things we didn't achieve (by design) with EOS/DPM & CASTOR
  - share the disk storage and namespace implementation
  - have a large (non-HEP) community product

Diamond R&D – Why a BUNLE?

- idea is to provide building blocks to customize storage infrastructure
  - few examples
    - XRootD / EOS Beryl today's building blocks
    - Scalable Meta Data Server (Namespace) with a parallel Query Engine
    - Scalable Storage System with File & Directory Interface based on common object storage interfaces (CEPH is placeholder)
    - XRootD + HTTPS interface for secure WAN access to object stores, local and network filesystems
    - POSIX FUSE interface with strong security
  - large community product

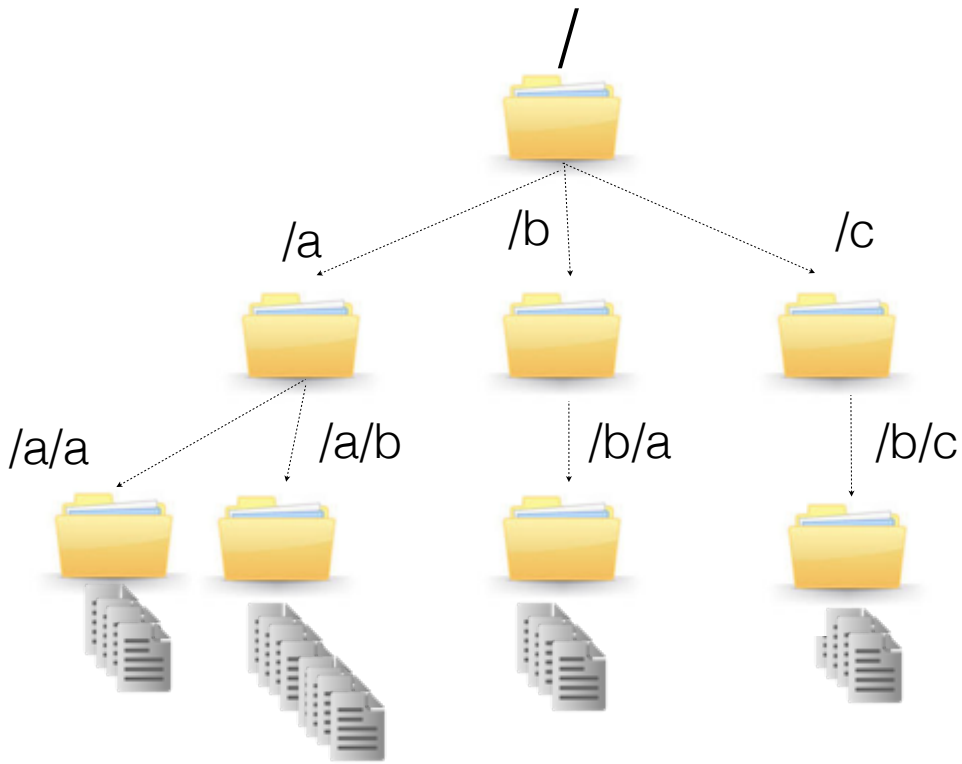
Scalable Namespace & Storage

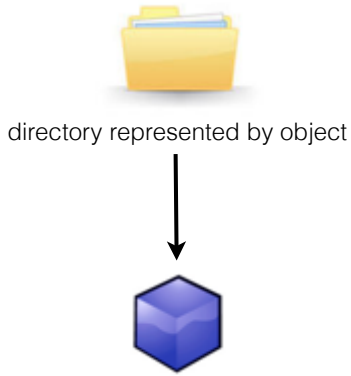
Infinity ∞

Diamond R&D

Infinity ∞

- trivial idea: store a namespace in a scalable object store
  - we can represent data in a *hierarchical structure* using directories and files and we *don't need* to group an infinite amount of files into a single directory
  - each *file* is a *change-log entry* in a directory object
  - each *directory* is represented as an *object* in an object store as a change-log file
    - these change-logs require compacting after many create/delete operations
    - a change-log file is perfect to cache remotely: if file size changed fetch the appended piece, if file size shrinks copy the whole file





dir.attributes

owner	perm	xattr
root root	xyz	user.x sys.y

file changelog

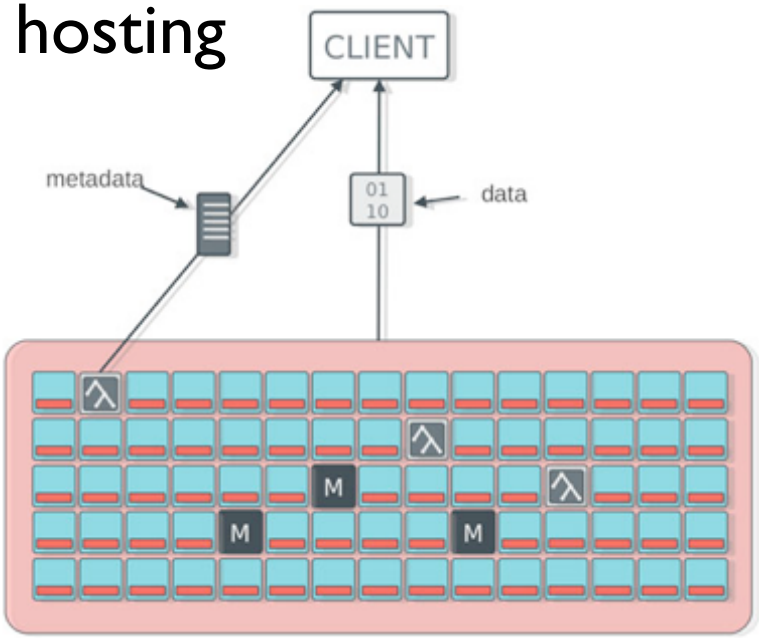
+ file1.root
+ file2.root
+ file3.root
+ file4.root
- file1.root
- file 2.root

An existing Object Store ...

Diamond R&D Scalable Object Store/Namespace using CEPH

- ceph is an open source implementation of an object store providing features like *dynamic resizing, self-healing, guaranteed consistency, low read latency, async object IO, extended attributes + key-value map per object, object notifications*

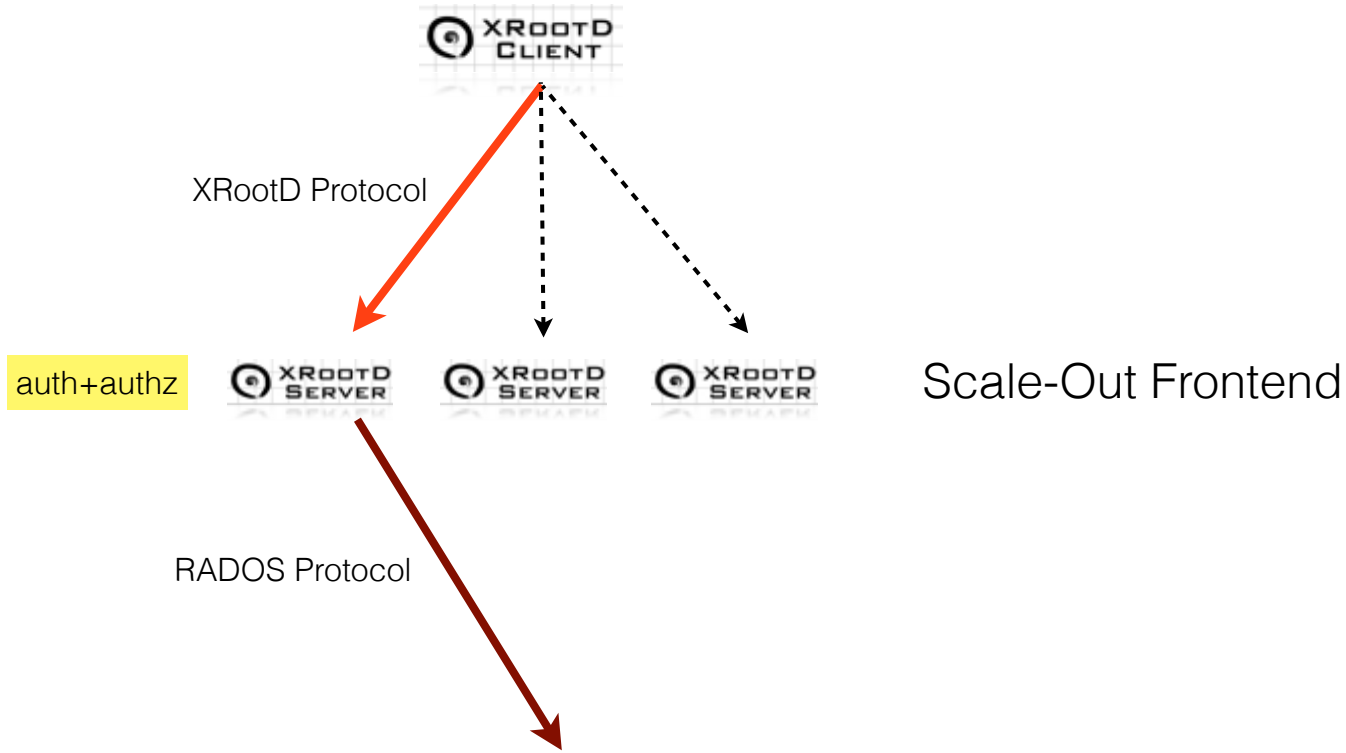
- IT-DSS provides now a ceph (rados) object store **service** with 1 PB capacity [x3] (~50 nodes) - initially for VM hosting



Diamond R&D Scalable Object Store/Namespace using CEPH

- ceph provides a filesystem - it is based on the paradigm to trust each client e.g. there is no server side authorization
  - non-trivial to add server side authorization
  - FS has (still) performance/stability/feature issues
- prototyping XRootD plugin using the object store implementation RADOS adding Auth+Authz to provide namespace and file storage
  - as a CEPH gateway running on front-end machines
  - as a CEPH overlay gateway on CEPH OSDs
- RADOS low-latency read, high-latency write ~25-40ms if not IO limited on hard disks
  (transaction ACK when synced in OSD journal - CEPH@CERN ~25k wOPS)

GATEWAY SETUP



easy deployment

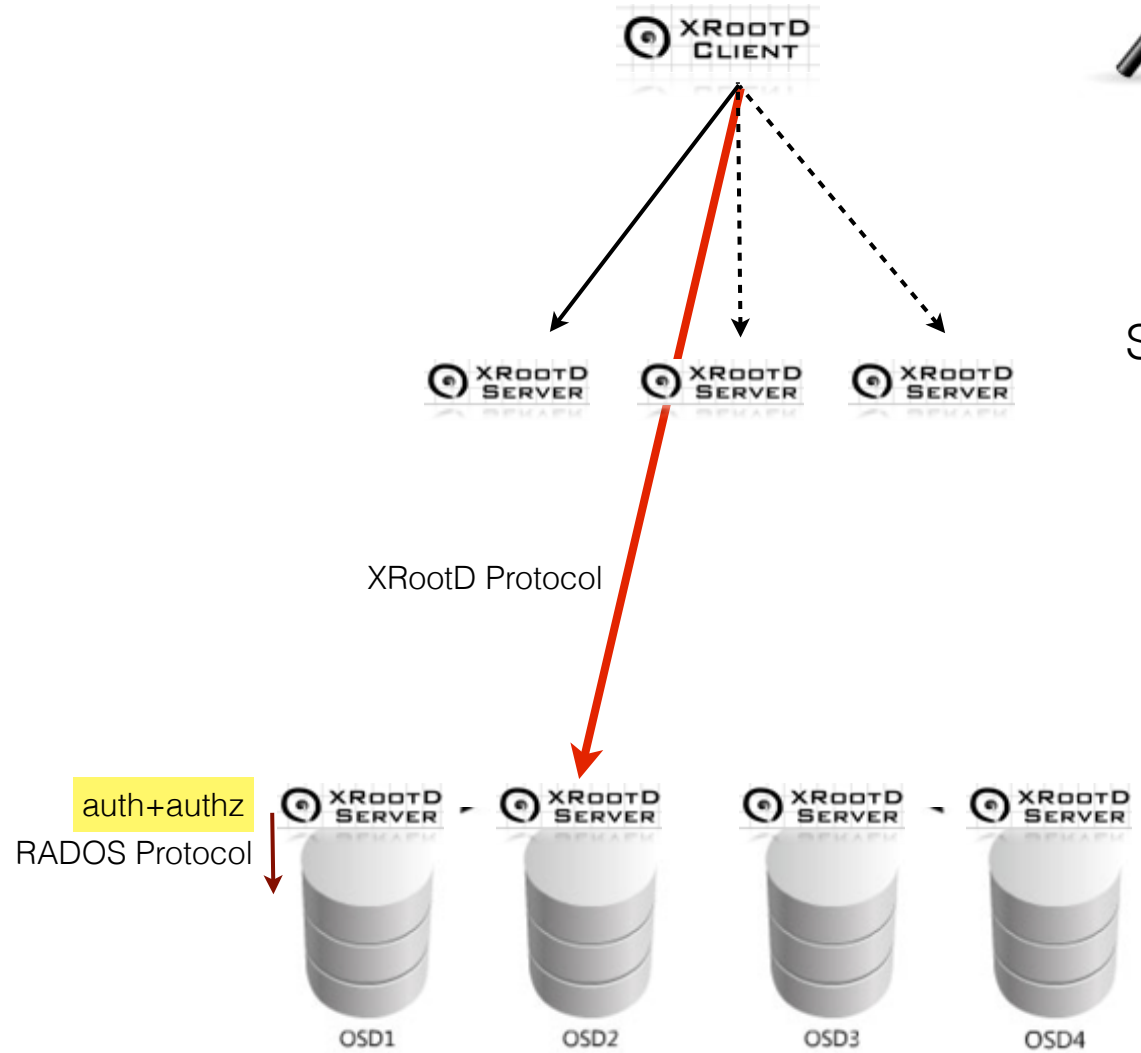
double IO traffic
double latency
loosing parallel IO

OVERLAY SETUP

Gateway on CEPH OSD

optimized IO

more complex deployment



Scale-Out Redirector Frontend

using RADOS API to compute location of a named object

ceph

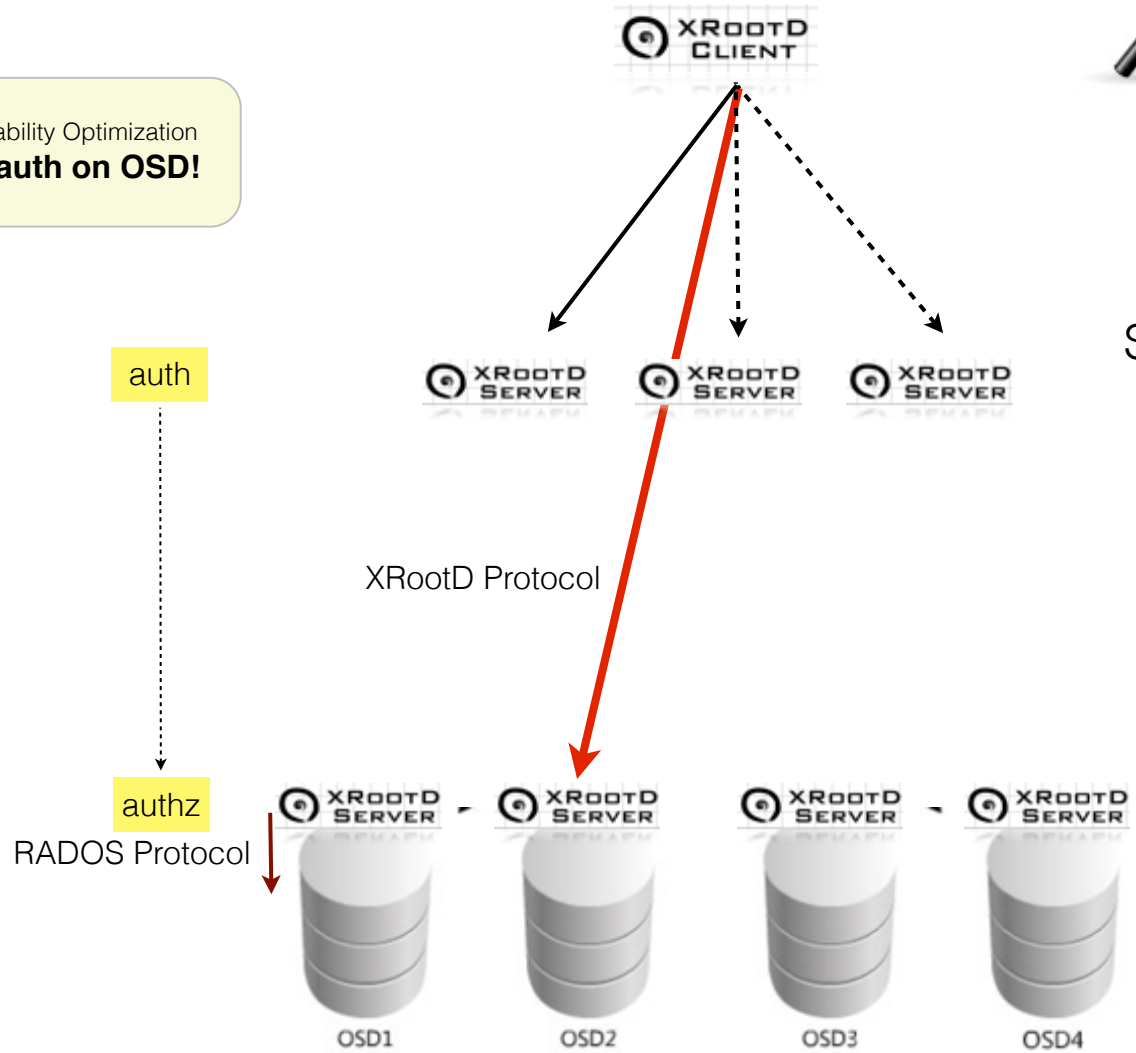
OVERLAY SETUP

Gateway on CEPH OSD

optimized IO

more complex deployment

Scalability Optimization
No auth on OSD!



Scale-Out Redirector Frontend

using RADOS API to compute location of a named object

ceph

Diamond R&D - Some Components ...

XRootD Auth Change Plugin

Plugin uses XRootD Authentication structure to set a per thread filesystem user/group ID.
Allows an XRootD to act like an NFS server storing and serving files with the uid/gid of the client connected e.g. can use the permission system of a local filesystem => allows AFS-like volume storage!

libRadosFS

IO library abstracting change-logfile object-based directories and object-based files

XRootD libRadosOSS

XRootD OSS plugin using libRadosFS to interface XRootD to the RADOS based pseudo-filesystem

XRootD libRadosCMS

XRootD CMS plugin to locate the primary location of RADOS objects for front-end redirection

XRootD libCephfsOSS

XRootD OSS plugin using libcephfs to interface XRootD to the CEPH Filesystem without a local mount

XRootD XrdCl Plugin

Plugin-in implementation allowing user transparent data management and parallel IO

Infrastructure-aware File Scheduling - UNITY

Server-side implementation as OFS plugin - client side as XrdCl plugin

FUSE IO & XrdCl Async/Cache Plugin

Refactor ll-FUSE implementation and XrdCl plugins to provide fully asynchronous file & dir operations and file & meta data caching

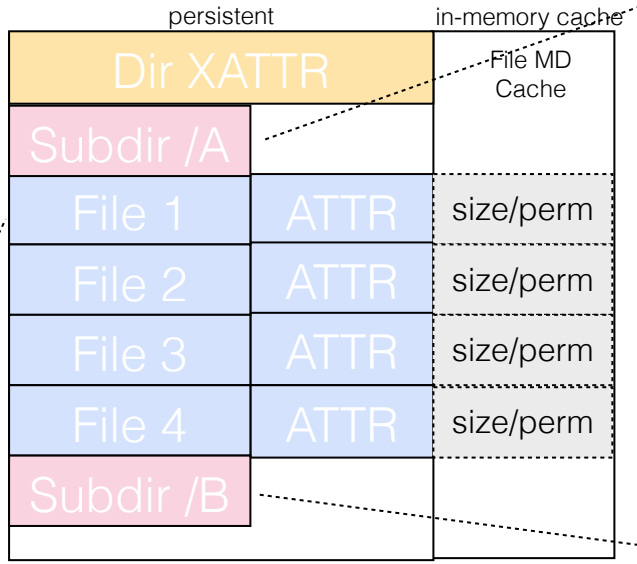
Diamond R&D - RadosFS ...

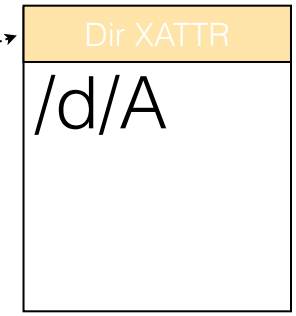
RadosFS

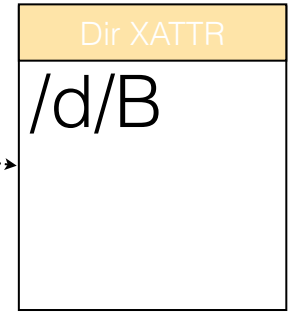
Optimizations:

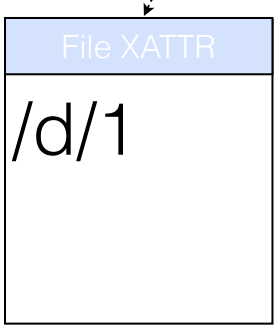
- File Access does not involve Directories
- Files are named objects with full path
- File location computed by object name
- Directories cache 'ls -la' related meta data
- Directories store query related attributes
- Renaming however is expensive (copy)

Directory Object /d









3 type of attributes:

<key,val> on dir object
(attributes related to the directory)

<key,val> on file object
(attributes related to the file)

<key,val> on entry in dir object
(query attributes related to files)

Diamond R&D - RadosFS - Current State

- File, Directory implementation complete
- XATTR on file and directories implemented
- ATTR on directory entry in progress
- 'ls -l' meta-data cache optimization in progress
- FSCK tool available
- Directory Object Compaction prototype implemented
- Query Engine not implemented yet
- Quota Accounting & ACLs not implemented yet

Tests so far:
- Prototype setup as **ATLAS small log-file storage** (few initial tests done)
- Test with Single-Core VM in OpenStack with **RadosOSS-XRootD** and CEPH service => [4k files] **creation 300 Hz open-read 1kHz**

- Tested 250 x 3 Mio objects in CEPH

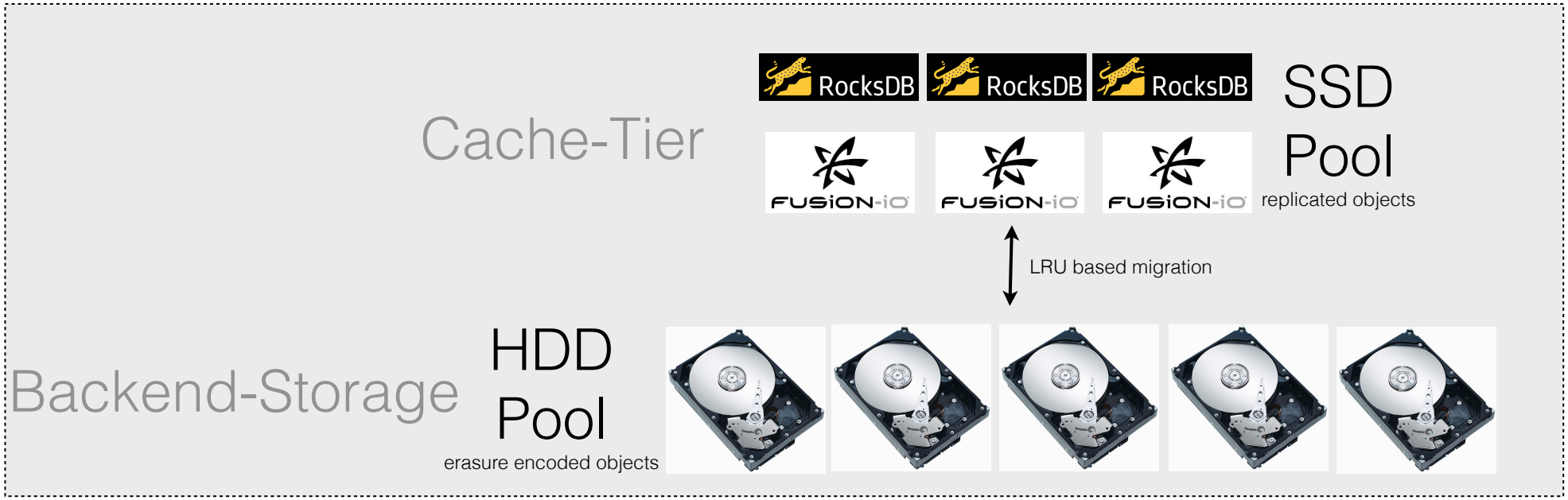
Diamond R&D - CEPH - Features

Current version of CEPH provides

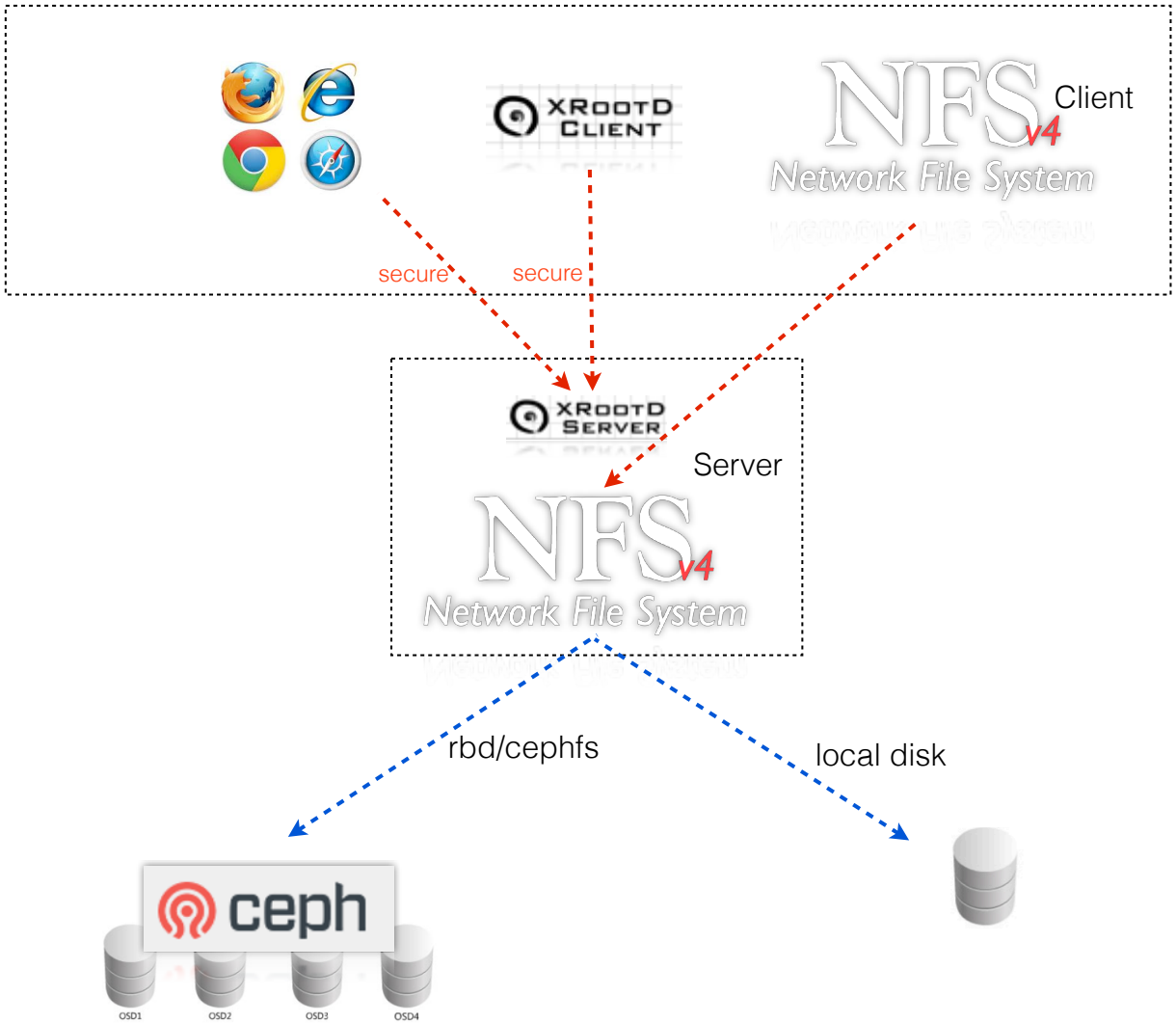
- object storage (rados)
- S3 storage (radosgw)
- rados block device (rbd) [VM hosting - ephemeral VM storage]
- 100% posix filesystem (CephFS) [main limitation: clients are trusted]

Next versions of CEPH (Firefly/Giant) contains

- storage tiering
- erasure encoding (currently alpha state)
- key-value store backend

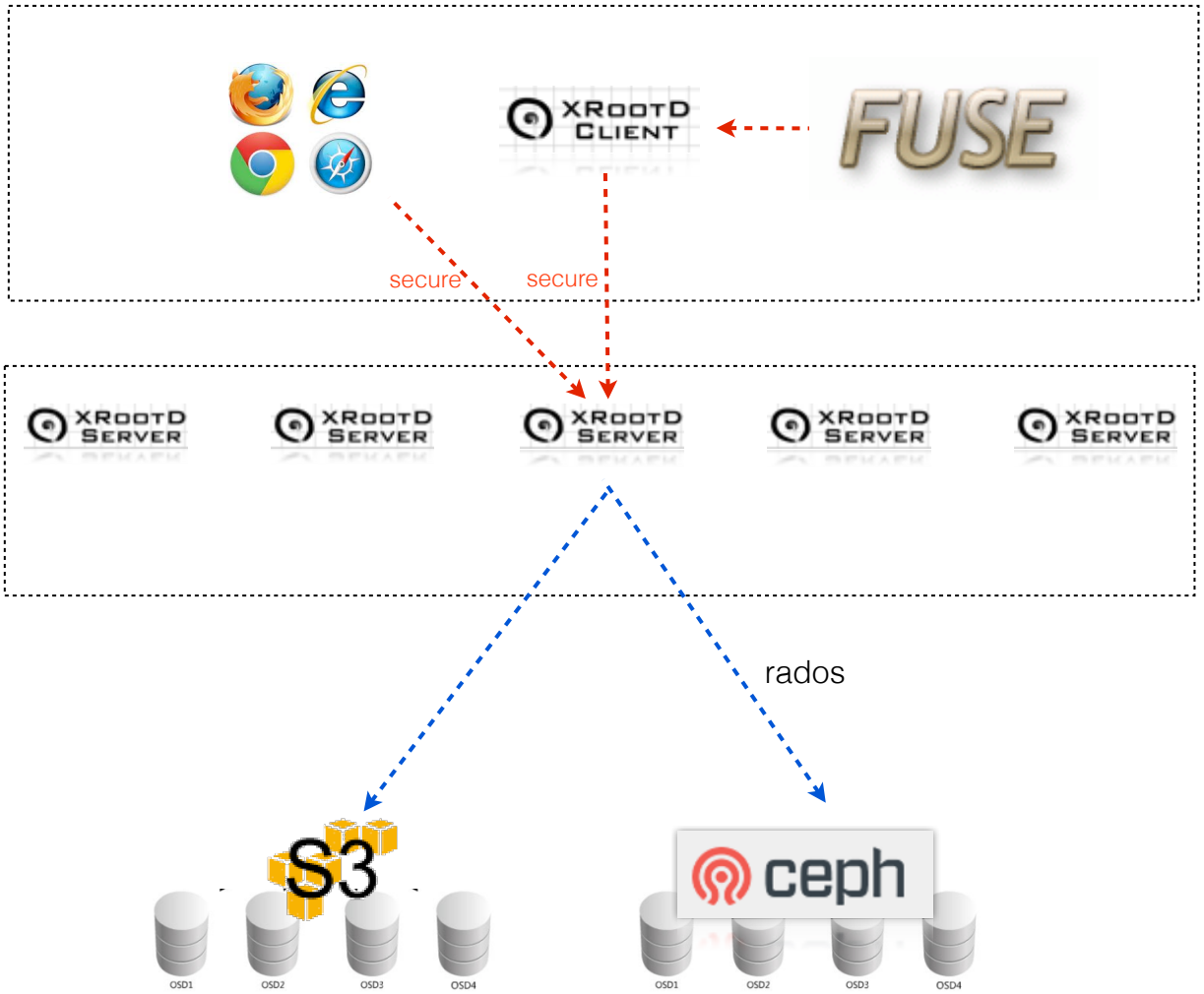


NFS Service with WAN & WebAccess



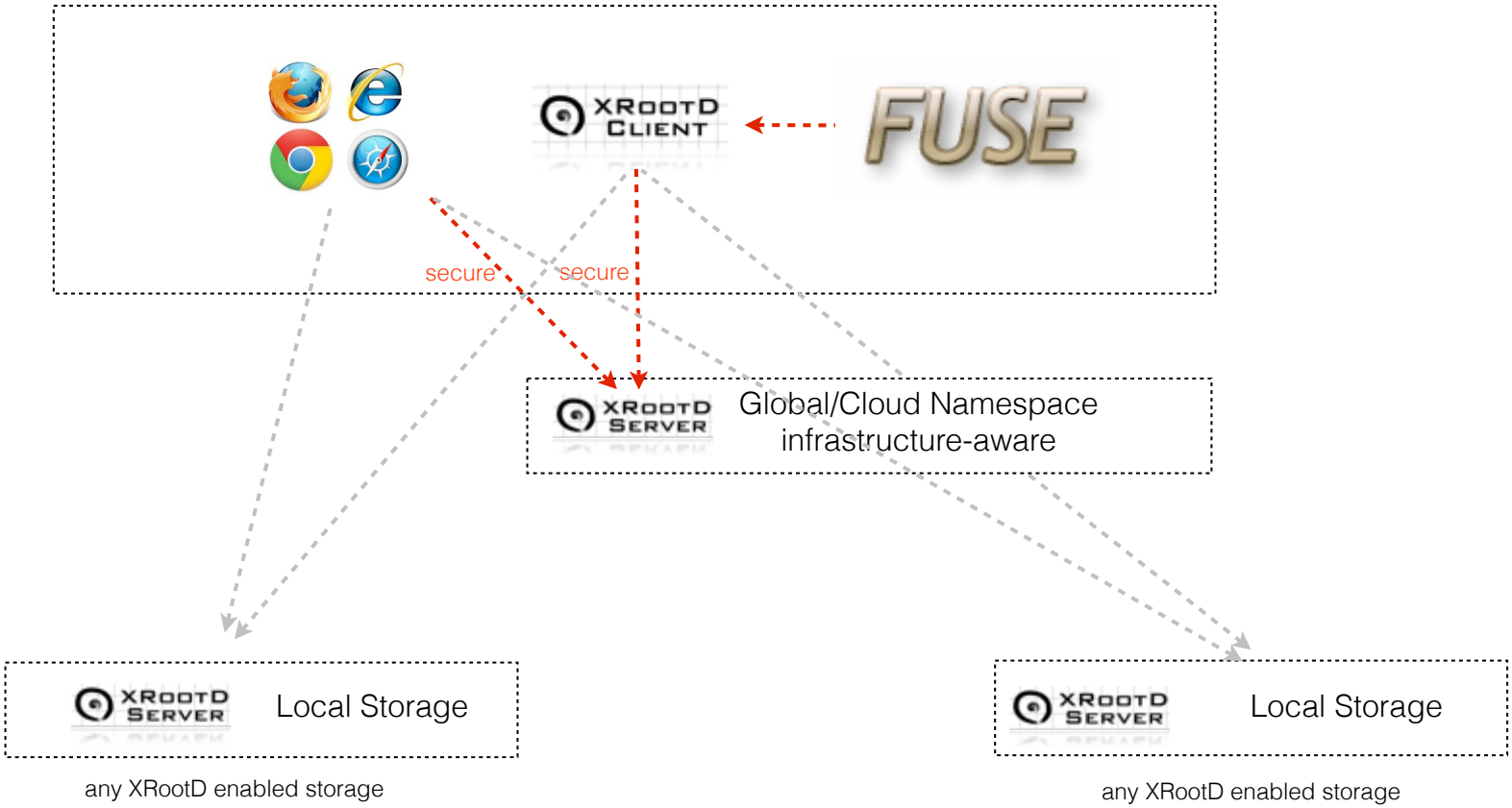
Diamond R&D – What can we build … another example

A Namespace for an Experiment Catalog or the successors of EOS and CASTOR



Diamond R&D – What can we build … 3rd example

A secure global or cloud data management and storage system

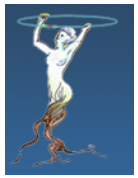


DIAMOND VST

"one plugin to rule them all ..."

DIAMOND Bundle

- ***XrdCl*** *IO Plugin*



root

FUSE

/global

xrdcopy

XrdCl IF

open hook

read/write hook

close hook

XrdCl Impl

geo-aware
policy-aware

VST

VST

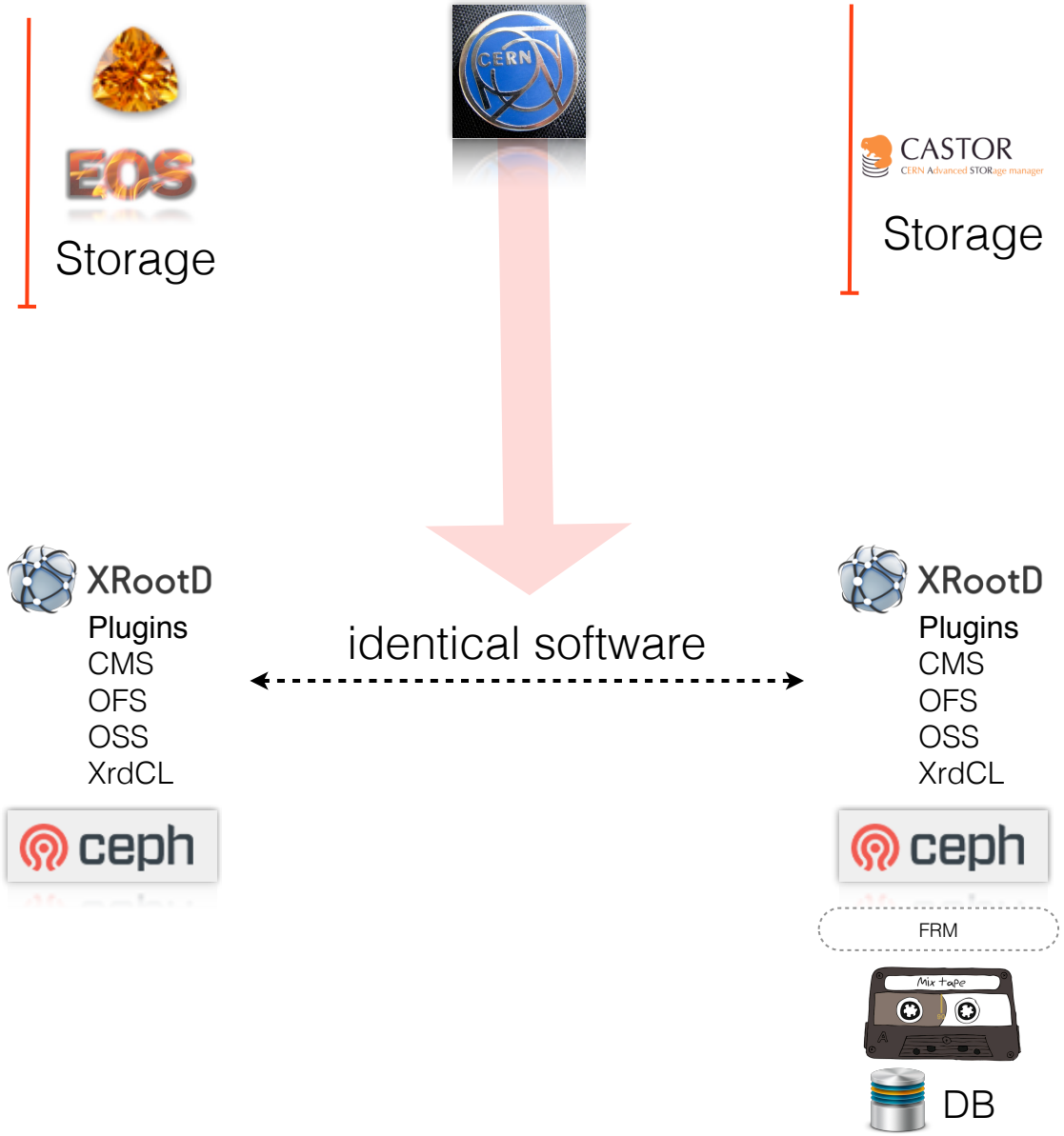
local XRootD enabled SE

VST

local XRootD enabled SE

- Gobal File Placement
- Global File Access
- UUID generation
- Checksumming
- Access Error Reporting

Diamond R&D – What we could aim in IT DSS as long term solution…



EOS Storage Bundle

CITRINE

XRootD

EOS

Storage

Self-contained & Simple HTTP/XRootD enabled Storage System

DIAMOND

XRootD

VST

ceph

Global DM

XRootD

ceph

radosFS (C) CERN

Storage

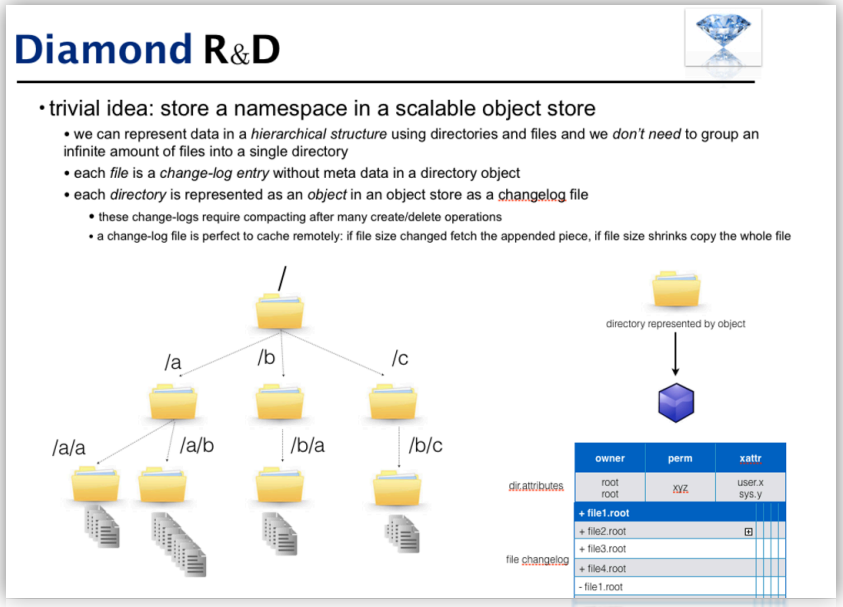
VM NFS

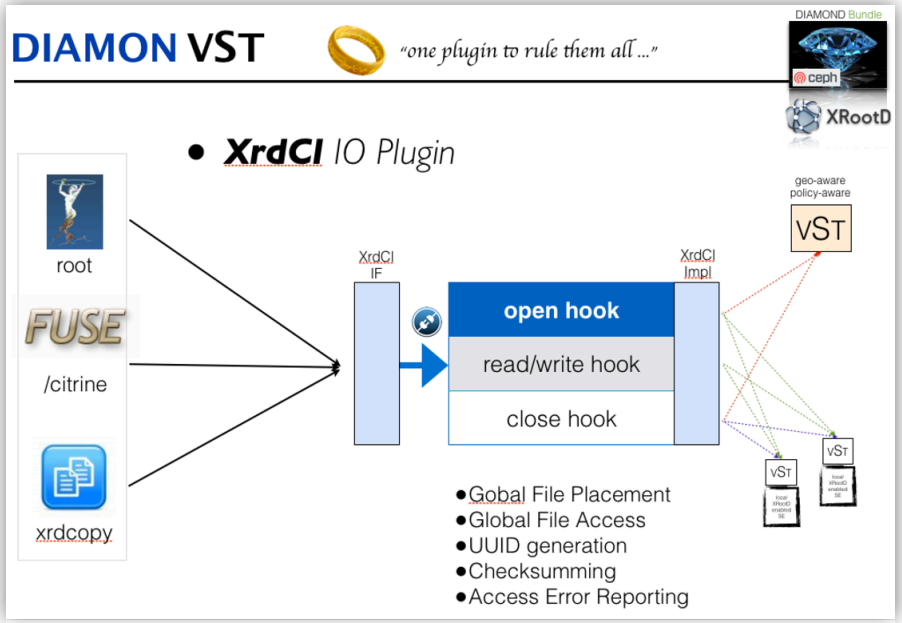
VM Hosting

Global Namespace & Metadata Catalog

Data Distribution Policies

Hybrid XRootD Storage





Diamond R&D – Timescale & Expectations

CEPH

- **very appealing** product as lowest storage layer - serves many use cases OpenStack/FileSystem/Object-Storage - Object Storage implementation very stable
- new interesting features like **Tiering & ErasureEC** probably **give 6-12 month** before production proven and ready
- integral part of OpenStack - **OpenSource**, backed by company Inktank
- CASTOR 2.1.15 will use parallel IO of CEPH to reach high-bandwidth tape streams

Caution: a fundamental concept of CEPH is to write small objects. A GB-sized file cannot be written as a single object. **CephFS** implements this chunking into 4 MB objects. CephFS is not yet considered production quality ... e.g. there is no FSCK check & repair utility and no support offered by Inktank. CEPH creates a strong coupling between hardware and failures. If the redundancy level is too low a node failure on the level of the default replication policy can make many or all files unavailable. Standard recommendation is three replica. Erasure EC will help improve this in the future. **CephFS** has to be run with XRootD in gateway mode e.g. CephFS is mounted on XRootD gateway nodes.

DIAMOND R&D

- **nothing decided** yet - still all R&D
- **start** performance & scalability **testing now** of RadosFS
- **afterwards** look at **implementation** of global DM tools
- intention to make a R&D package available soon for the Firefly release for interested people **'to play'** and gather some experience

Thank You