



Open Source Storage



Open Source Storage

andreas.Joachim.peters@cern.ch

Storage Systems for Big Data

Andreas-Joachim Peters
CERN - IT
Storage Group

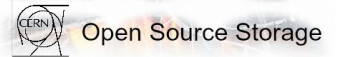
**WORKSHOP OF WLCG RESEARCH PROJECTS
FOR HL-LHC ERA EXPERIMENTS**



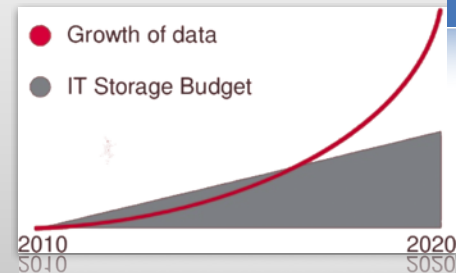
10 - 12 JULY 2017
SAINT-PETERSBURG, RUSSIA

Contents

- Reminder on Big Data technology
- EOS service at CERN
- Development & Service roadmap
- Federations & HI-LHC Future



" **Big Data** is a large volume of unstructured data which can not be handled by standard [database](#) management systems like [DBMS](#), [RDBMS](#) or [ORDBMS](#) "



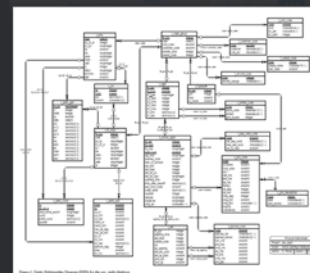
- most of data stored is **unstructured data** (photo, audio, video)



data for humans

best match
Scale-Out Storage

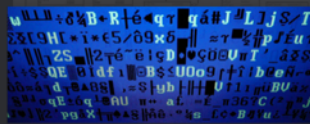
- small fraction of **structured data** (DBs, derived data, meta data)



machine data

mostly stored in
Scale-Up Storage

- **Raw Data**
(LHC)

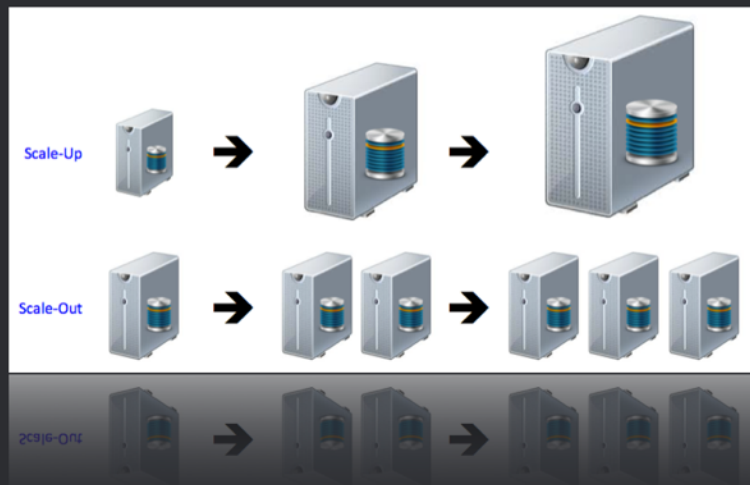
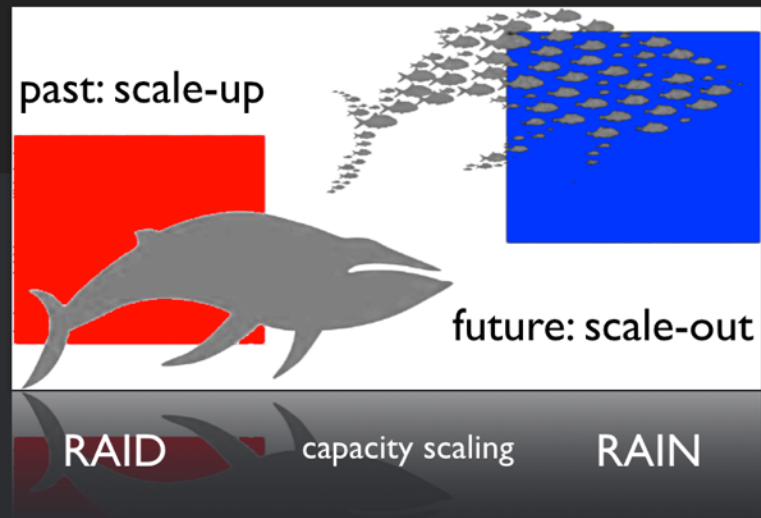


best match
Scale-Out Storage
Tiered Storage



'Big Data - The Solution'

From Scale-Up to Scale-Out Storage



- structured data with central view & strong consistency
- predictable access
- SPOF
- slow growth rate

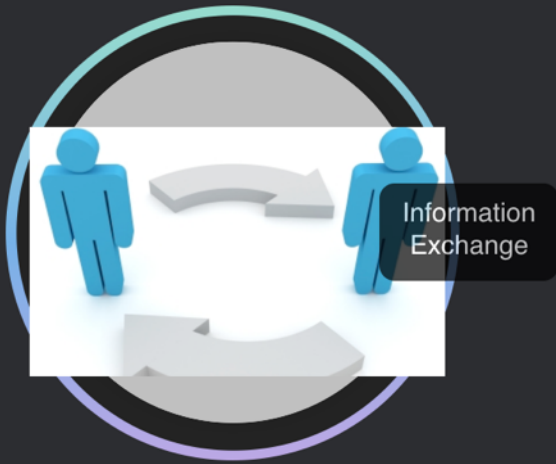
- best for unstructured data
- unpredictable growth rates
- MPOF resistant

▶ Big Data - Hyper Scale Computing

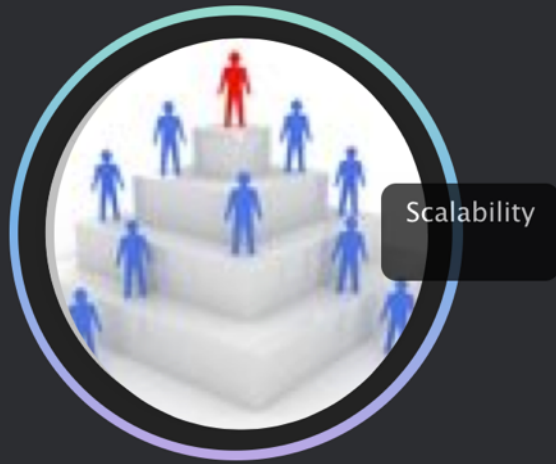
Amazon, Google, Facebook ...

- **server** and directly attached storage becomes **basic unit** to build clusters of thousands of nodes using commodity hardware
 - **no redundancy within a single node** necessary
- instead of many specialised applications small number of huge applications
 - **no enterprise IT environment**
- open source platforms for data & storage services
 - **CEPH**
 - **Swift**
 - **Cassandra**
 - **Riak**
 - **Hadoop**
 - ...
- requires **automation** of node deployment, **failure recovery**
- **Software defined Data Center/Storage**
 - **System On Chip Server** (e.g. HP Moonshot) - space efficient, energy efficient
 - **Disk embedded Server** (e.g. HGST, Seagate Kinetic)

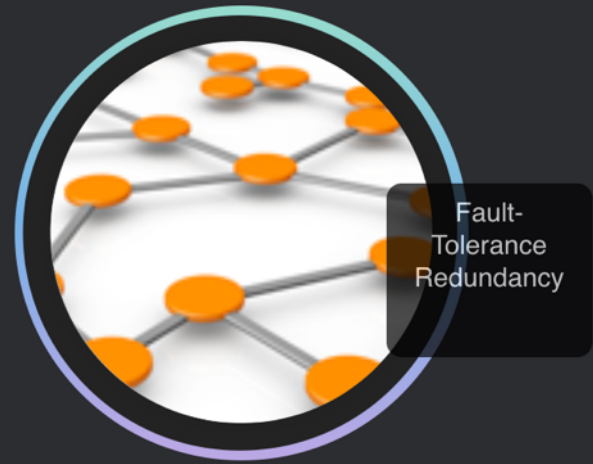
▶ Base ingredients of Big Data Storage & Hyper-scale Computing



How to get to data ...APIs ...



What to scale and how ...



How to avoid data loss ...

CERN Storage Infrastructure

Backup
Archive

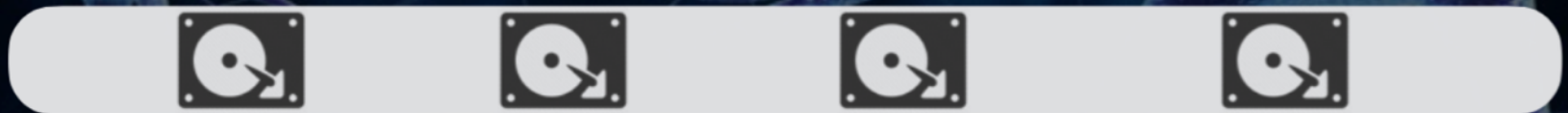
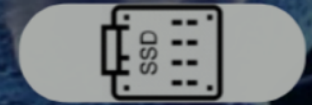
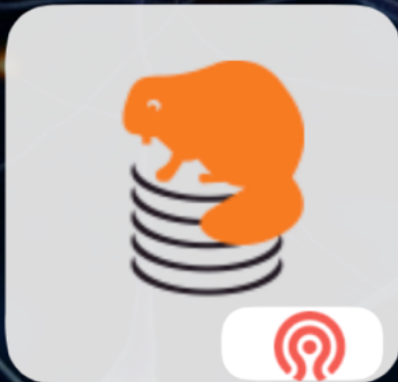
Big Data
Storage

Home
Directory

Software
Distribution

Shared
Storage

VM Block
Storage





Open Source Storage

first production service in 2011

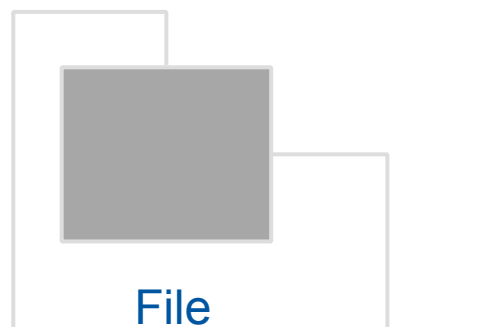


Open Source Storage

Technological Foundation

- Organic file storage system built on XRootD framework
- disk-only storage system on top of **JBODs** with replication & erasure encoding
- open source project in CERN IT storage group
 - released under **GPLv3**
- extremely simple architecture
 - one daemon and three plugins written in C++
- no relational DB backend, namespace is in-memory stored in a changelog file, no commercial product dependencies
- easily adaptable to requirements
- **CERN** code ownership
 - agile development style
- very short development & release cycles

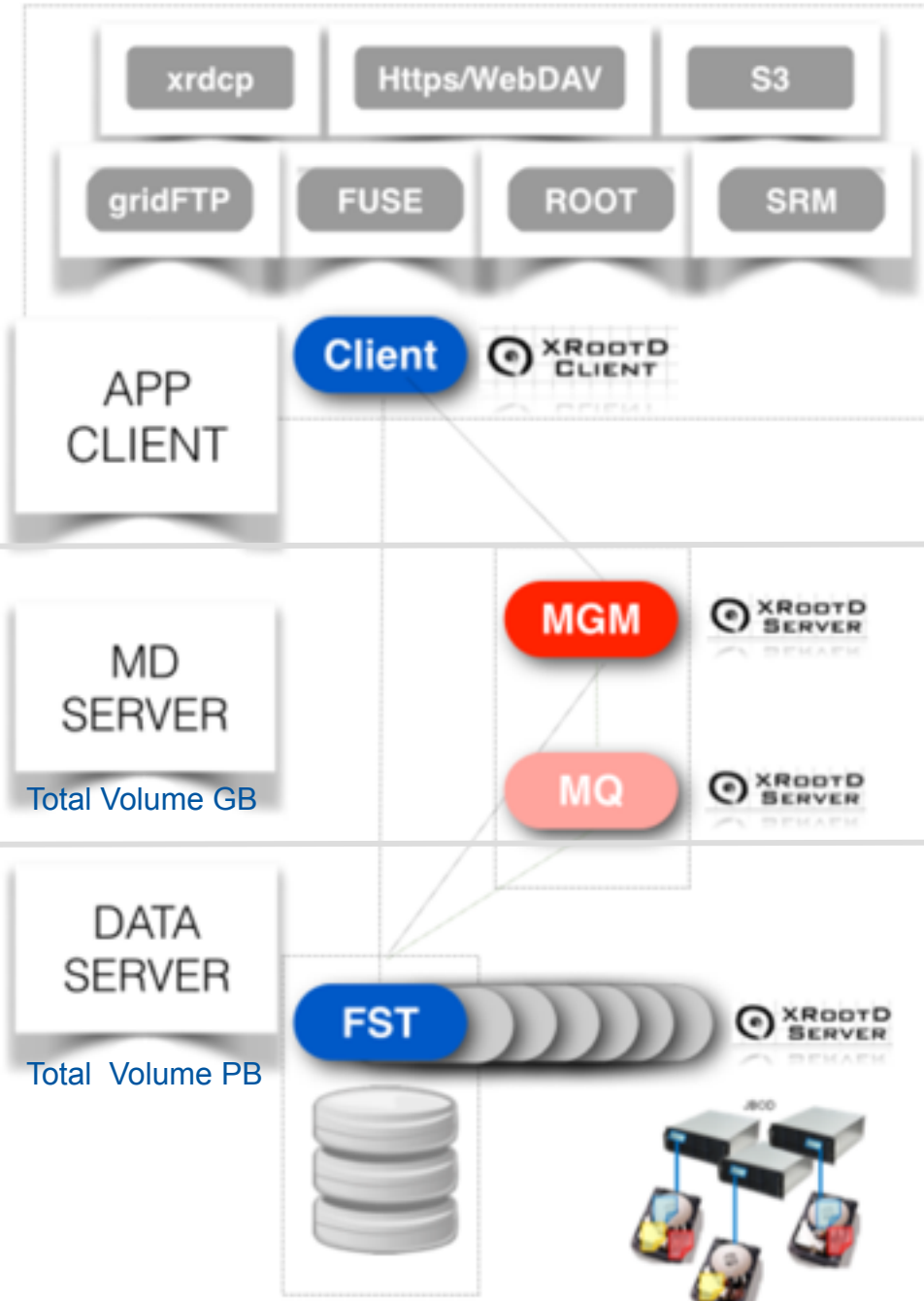
Architecture



Who owns it?
When was it created?



Contents



Components

Management Server

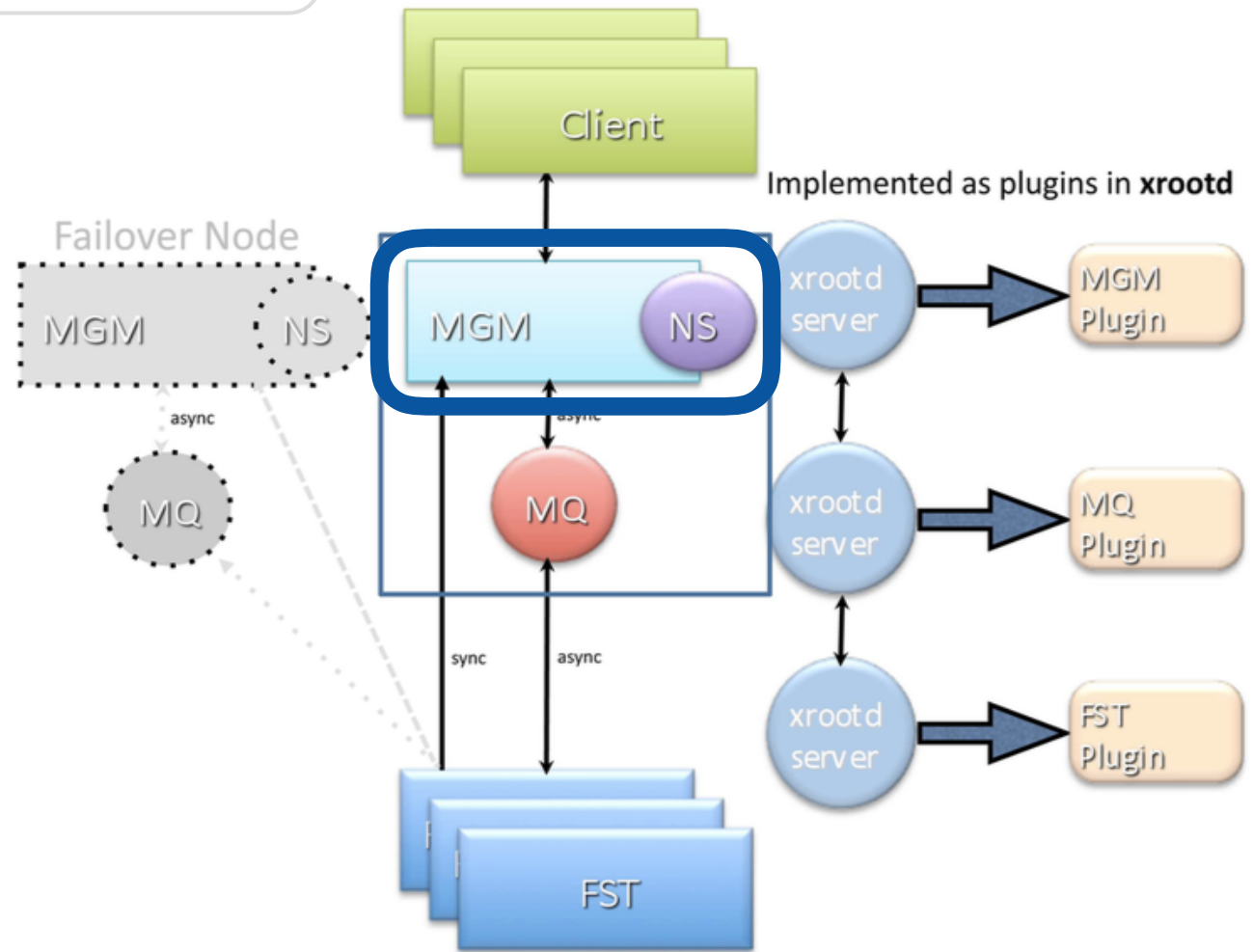
- Pluggable Namespace, Quota
- Strong Authentication
- Capability Engine
- File Placement
- File Location

Message Queue

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

File Storage

- File & File Meta Data Store
- Capability Authorization
- Check-summing & Verification
- Disk Error Detection (Scrubbing)



Components

Management Server

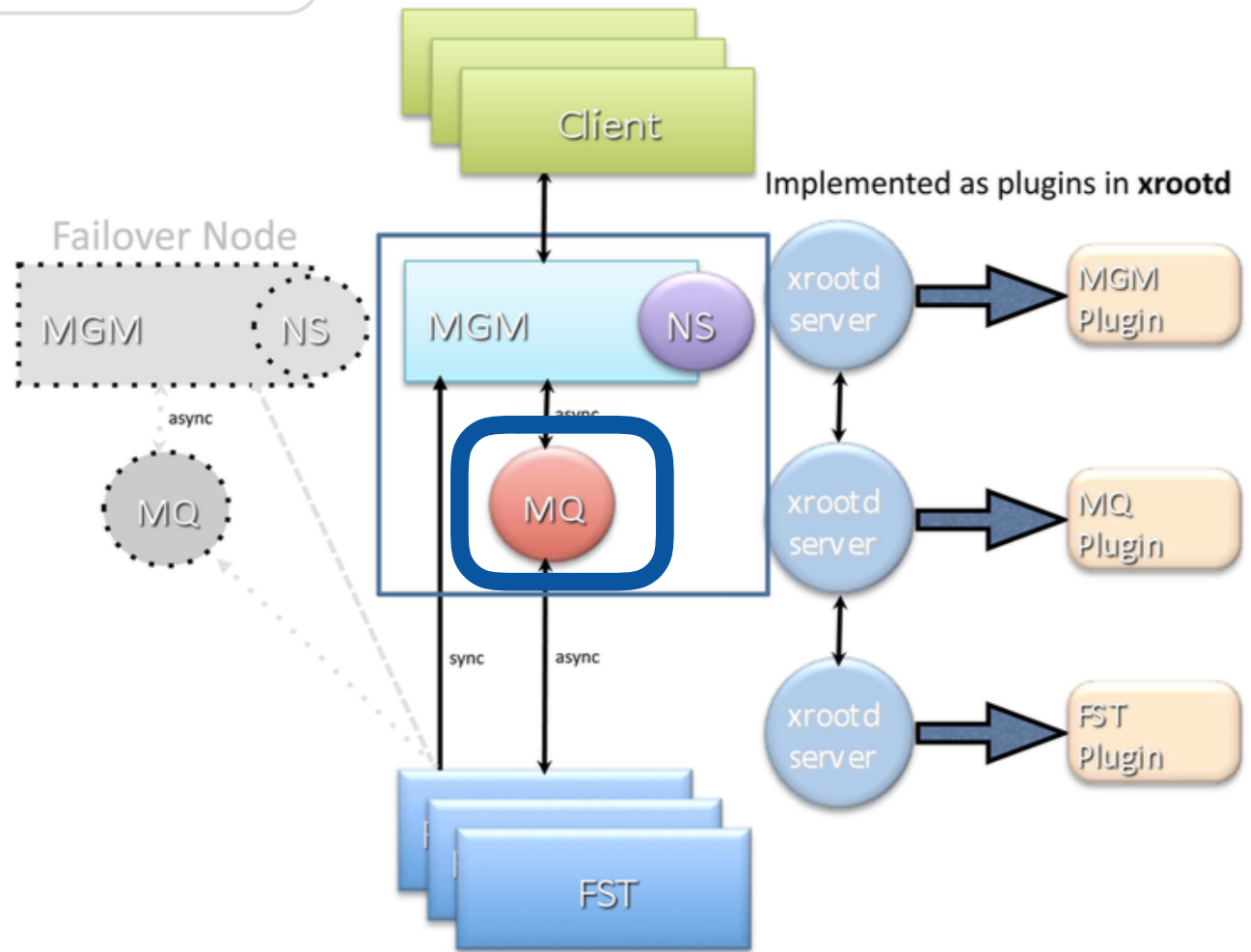
- Pluggable Namespace, Quota
- Strong Authentication
- Capability Engine
- File Placement
- File Location

Message Queue

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

File Storage

- File & File Meta Data Store
- Capability Authorization
- Check-summing & Verification
- Disk Error Detection (Scrubbing)

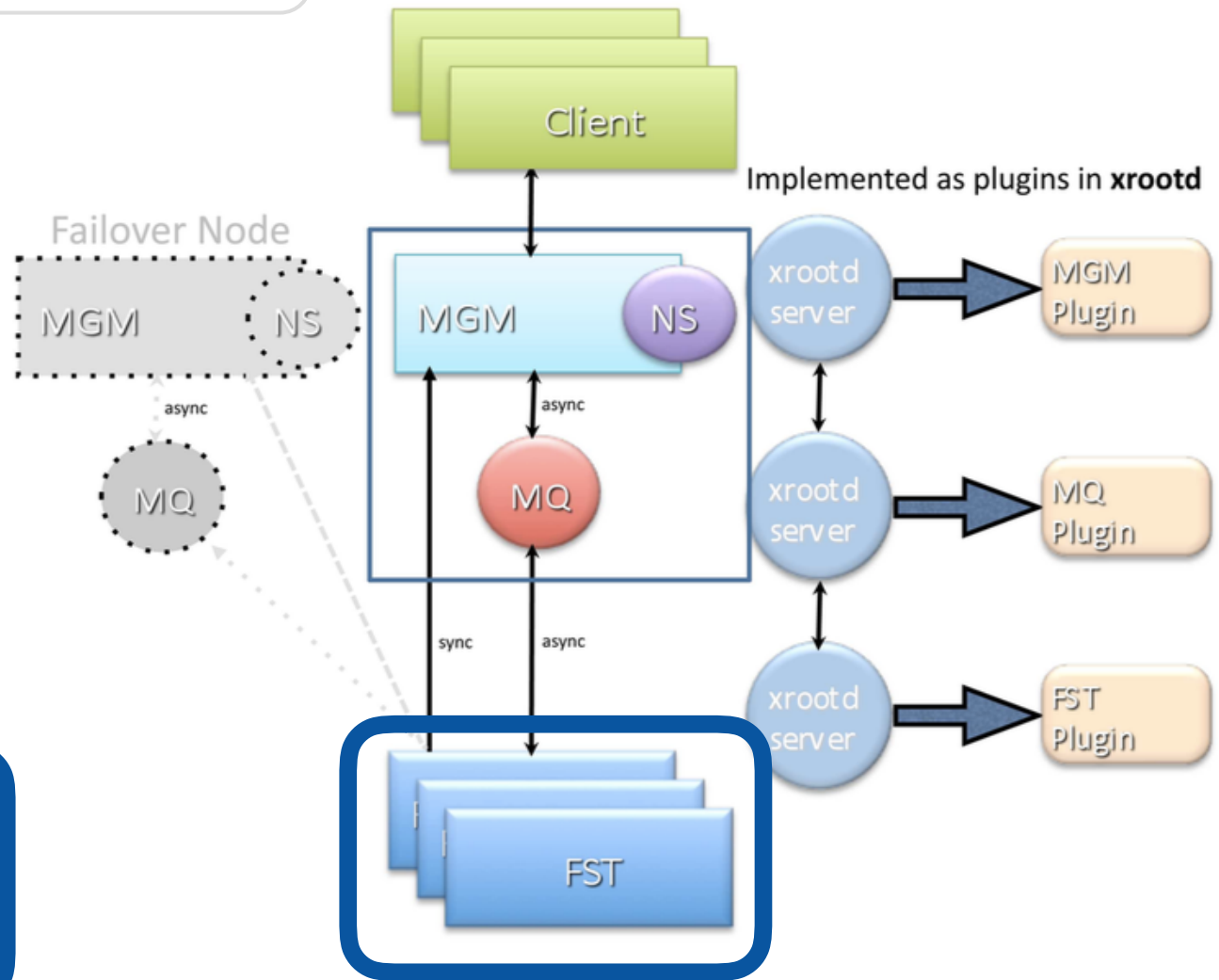


Components

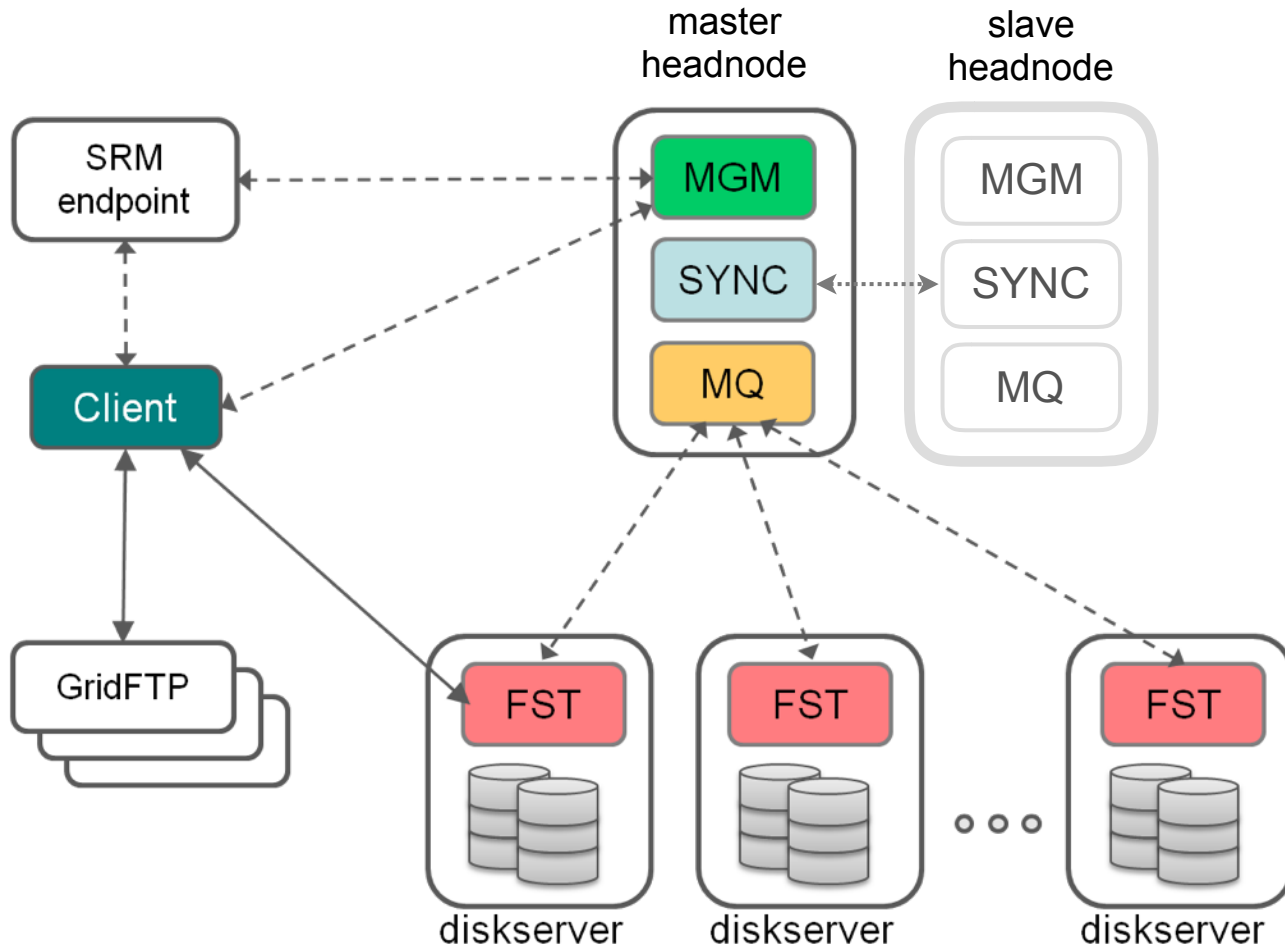
Management Server
Pluggable Namespace, Quota
Strong Authentication
Capability Engine
File Placement
File Location

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

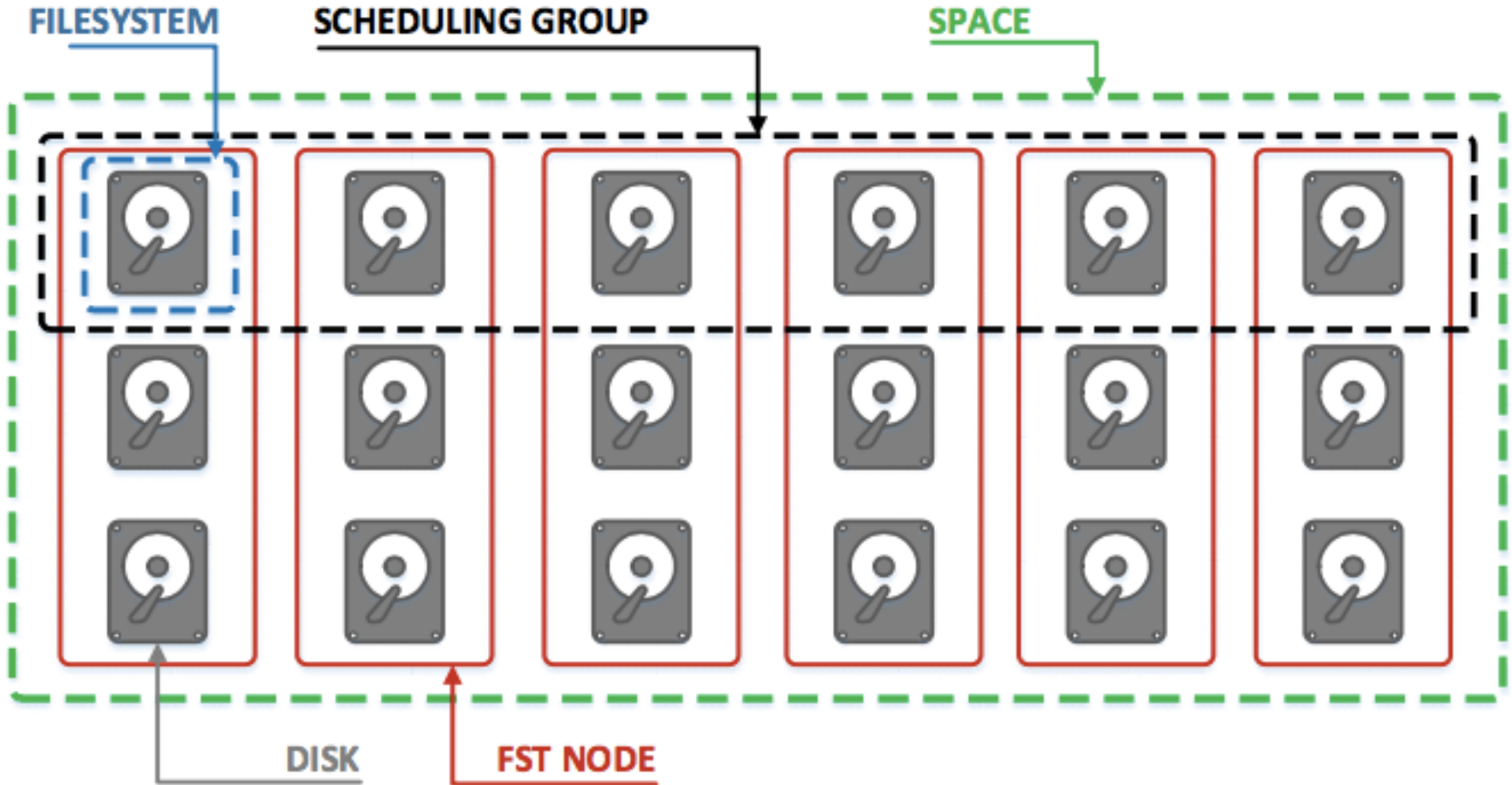
File Storage
File & File Meta Data Store
Capability Authorization
Check-summing & Verification
Disk Error Detection (Scrubbing)



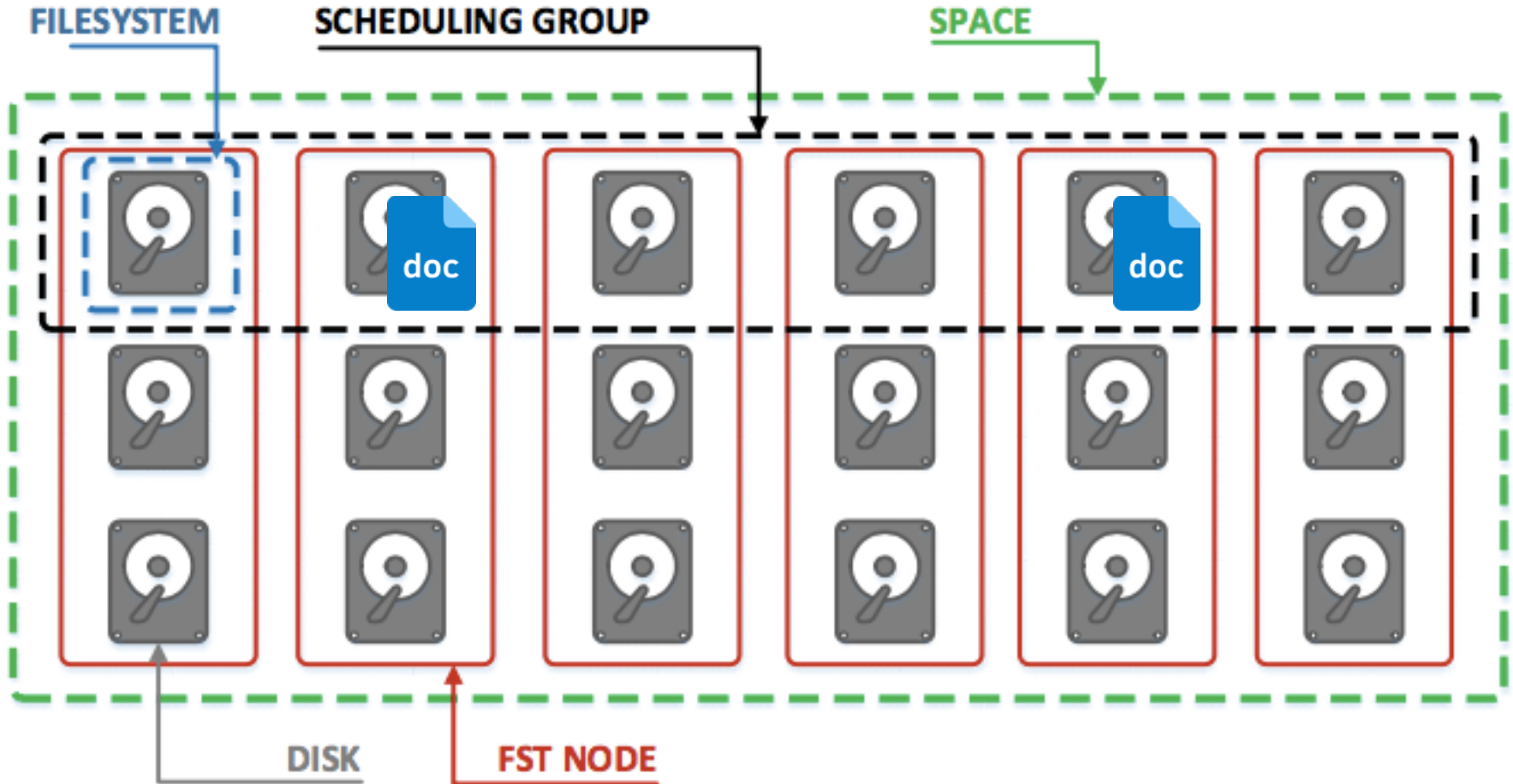
Deployment View



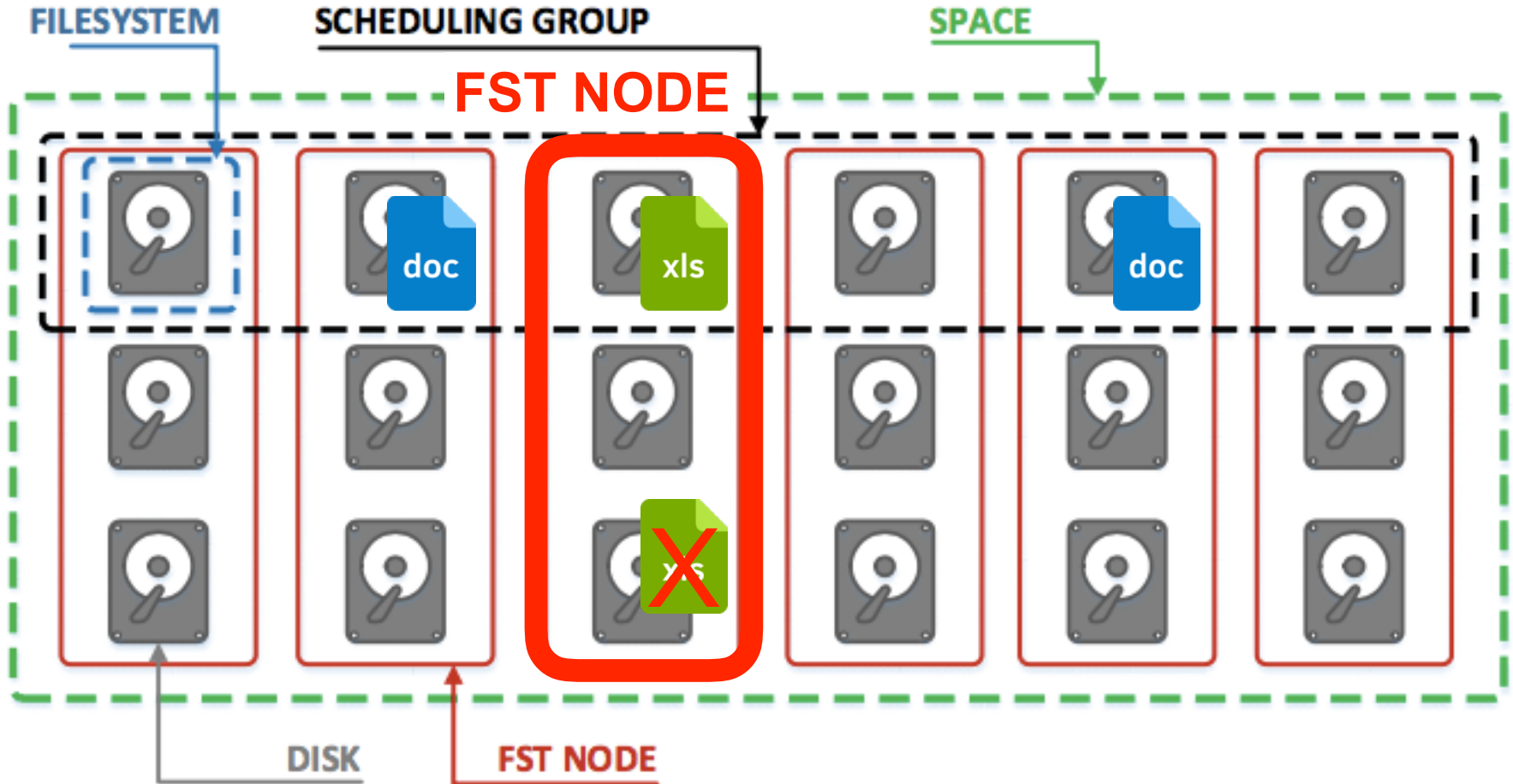
Data Placement



Data Placement



Data Placement



node failure = data unavailable

EOS Releases

named after gemstones



Beryl Aquamarine
V 0.3.X



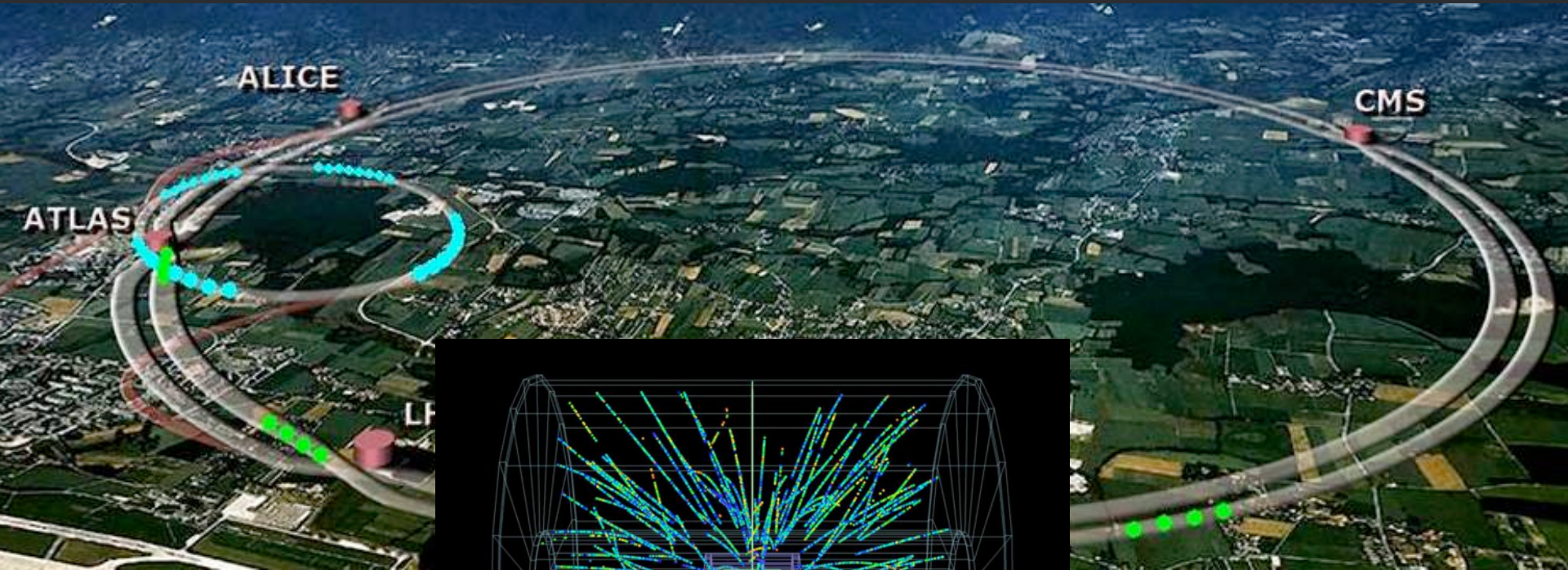
Citrine
V 4.X

XRootD V3 Server
IPV4
namespace in-memory
data on attached disks

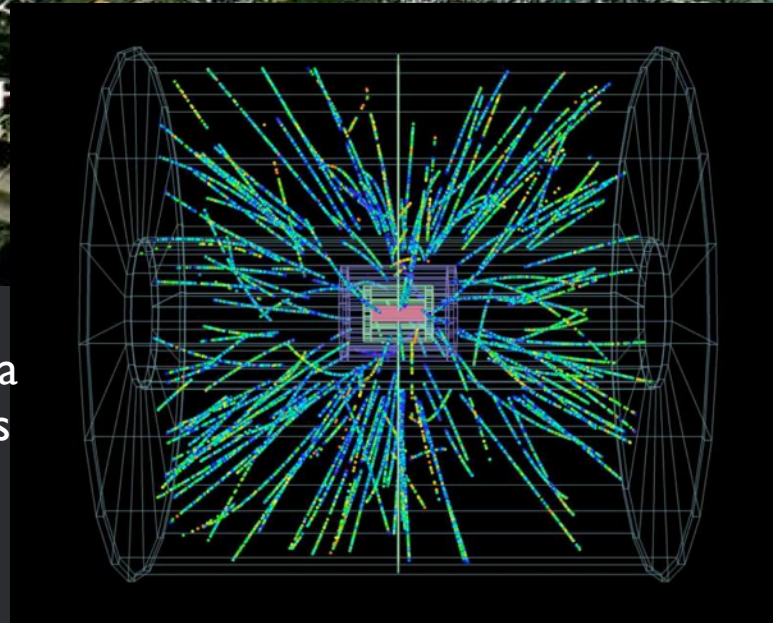
XRootD V4 Server
IPV6
plugins for meta
data & data persistency

scale-out data
scale-up meta-data

scale-out data
scale-up & scale-out meta-data



theoretical unfiltered data stream of LHC detectors
182 ZB/year



What happens at CERN storage?



EOS at CERN



 Open Source Storage

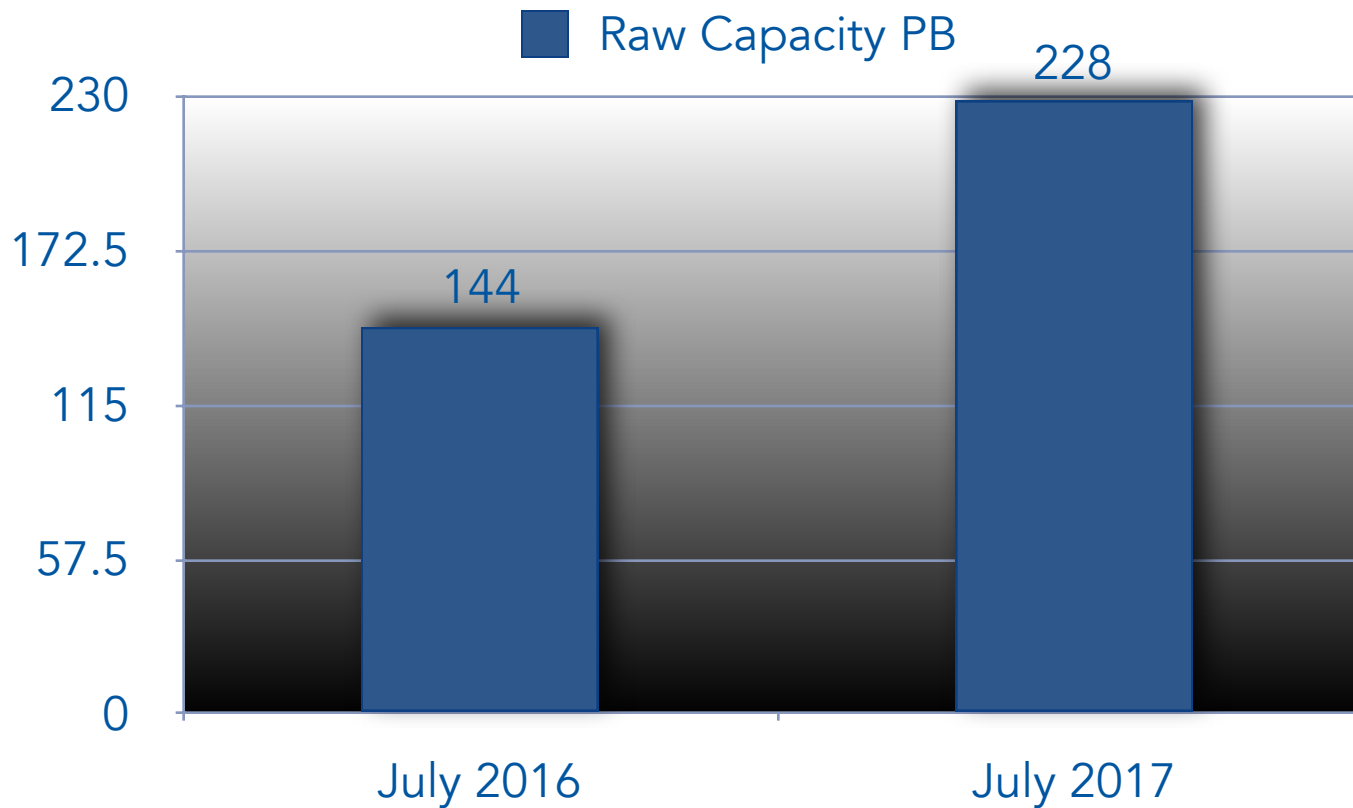


 Open Source Storage

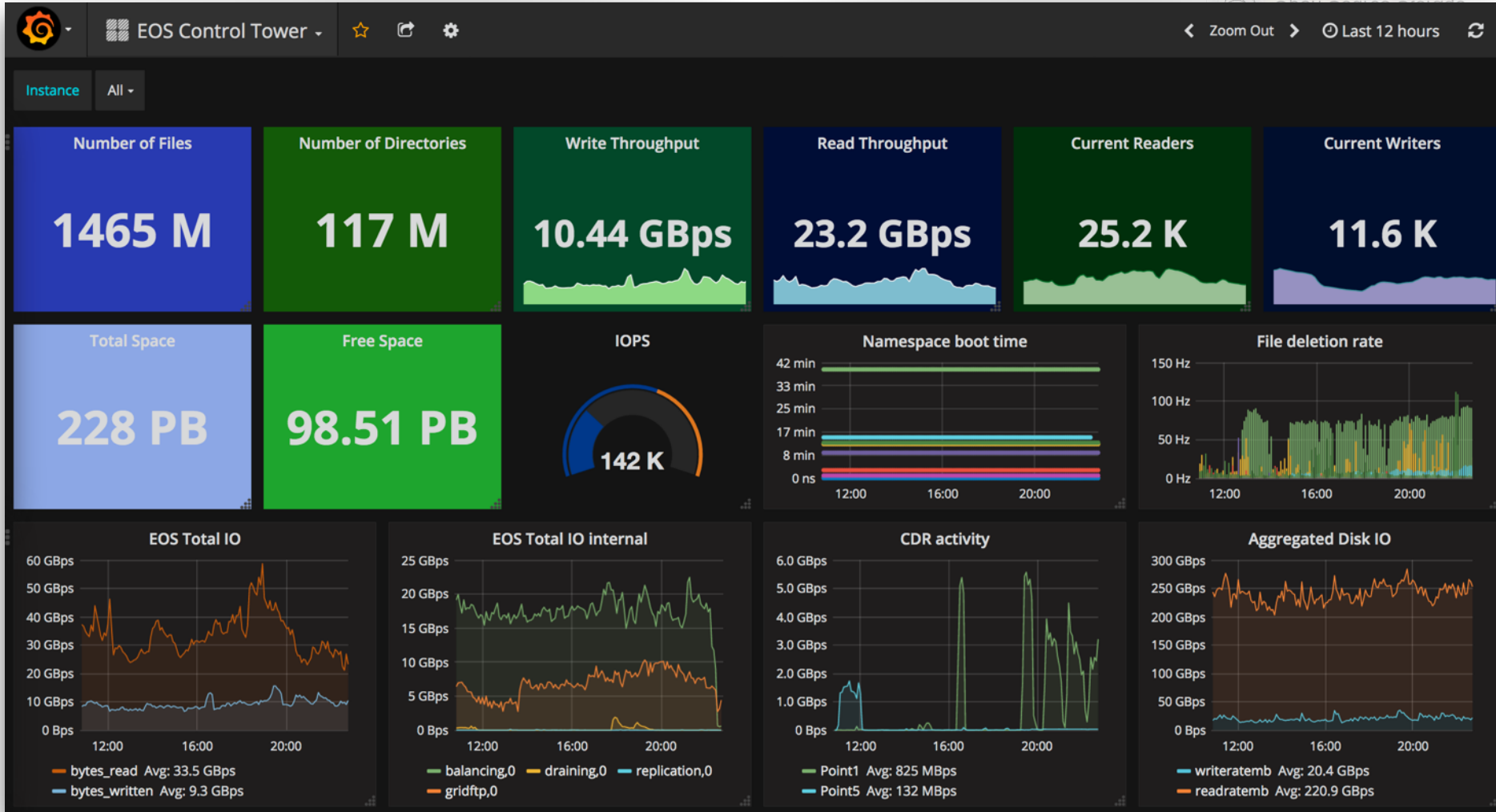
1 year averages: 38 GB/s read 12 GB/s write

1.2 Exabyte/year

378 Petabyte/year



EOS at CERN



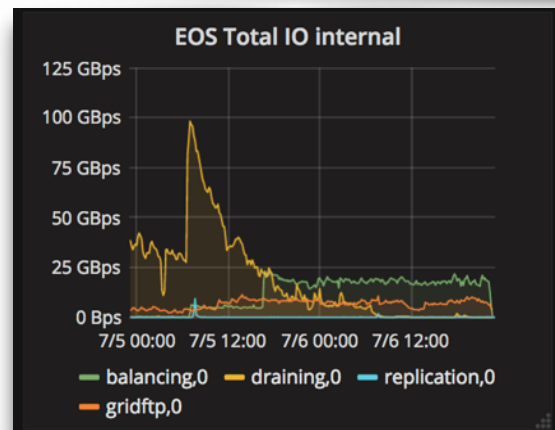
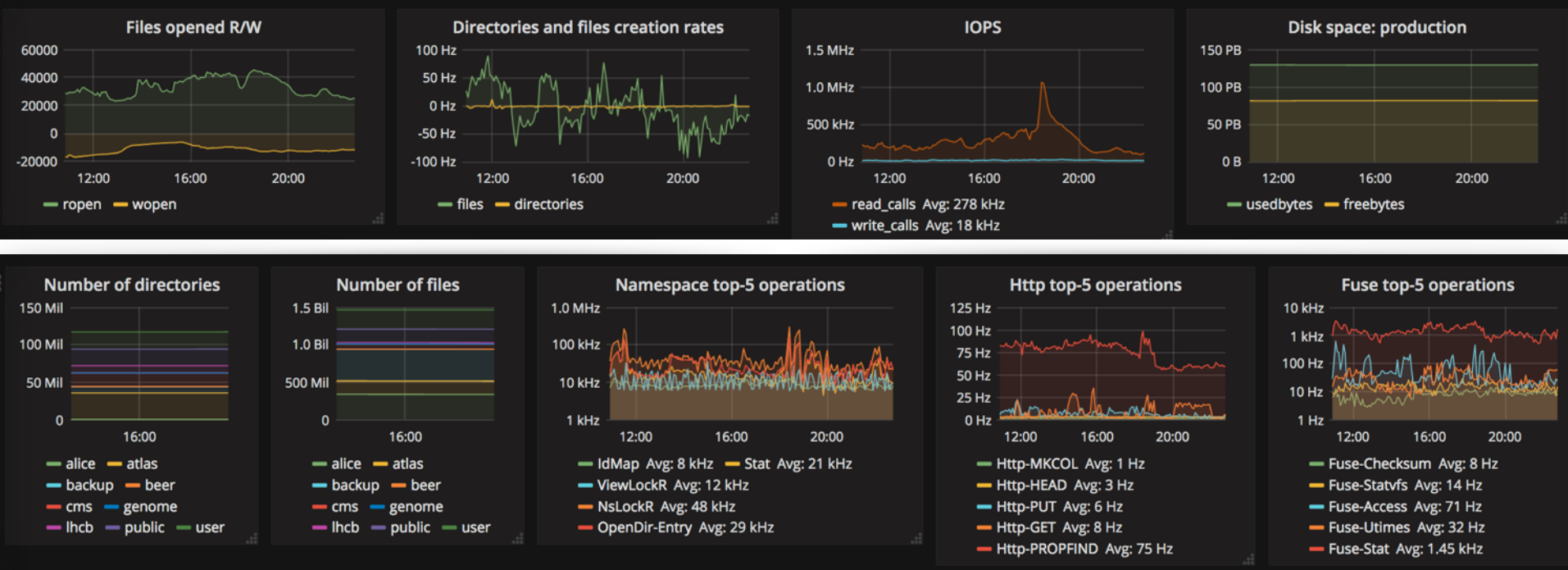
scrubbing 7.8 Exabyte/year

EOS at CERN



Open Source Storage

observed




Recently drained many PB
Draining peaked at 100 GB/s



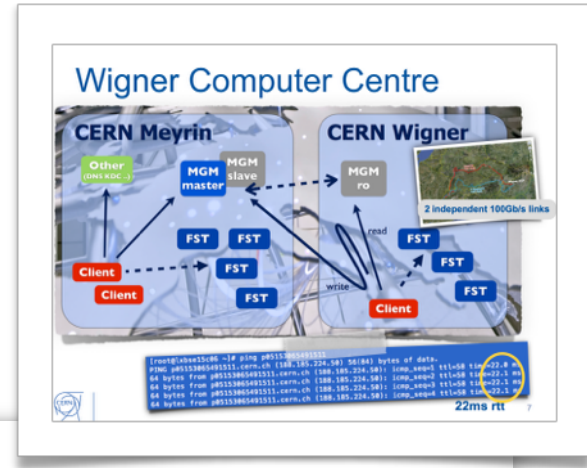
EOS CERN Instances



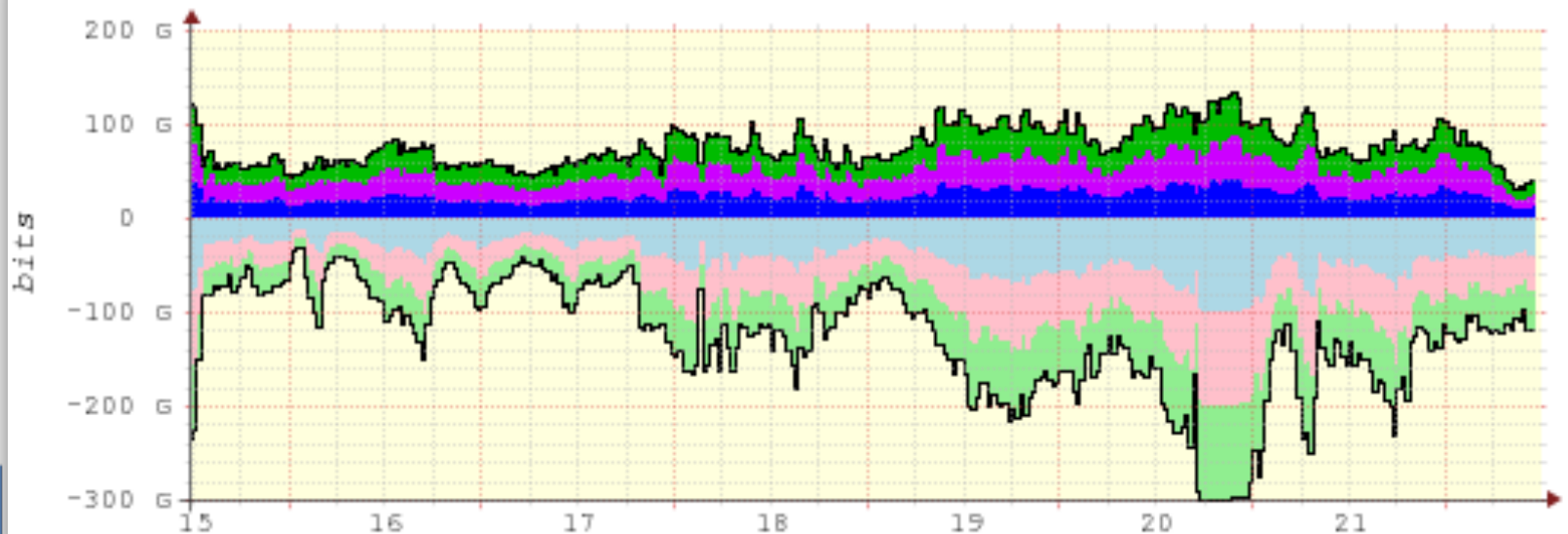
 Open Source Storage

 Open Source Storage

- Disk Storage for all LHC and physics data
 - and for CERNBox
- Deployment across two computer centres
 - CERN and Wigner RCP - 22ms RTT
 - Third link recently added




Total Traffic to/from Wigner



EOS CERN Instances



 Open Source Storage

 Open Source Storage

ALICE

ATLAS

CMS

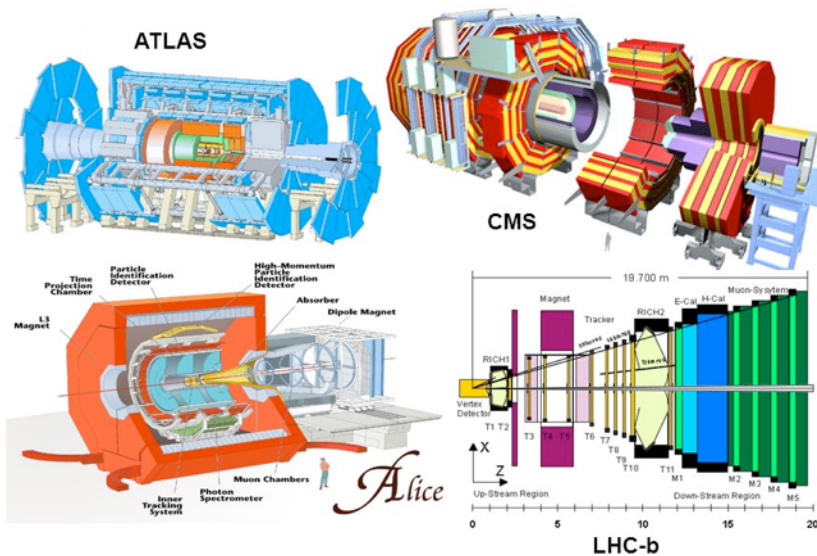
LHCB

PUBLIC

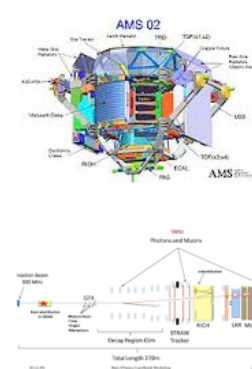
USER

Backup

LHC Experiments



non LHC Experiments



work spaces

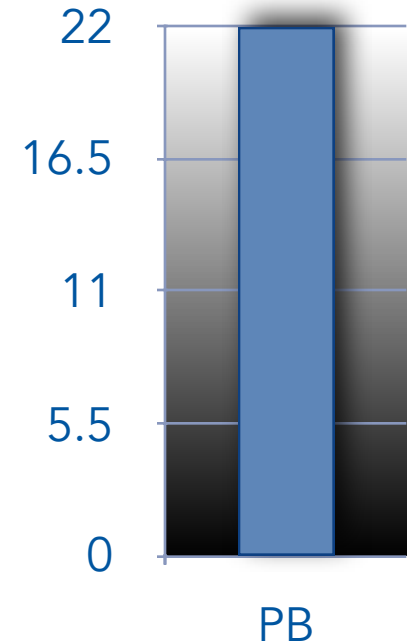
project spaces

Instance Challenges



ALICE

- largest namespace 340 M files 86k dirs
- simplest IO model - XRootD
- simplest user model - single GRID user
- pre-signed URLs
- thousands of LAN/WAN connections
- no quota
- dominated by GRID analysis activity
 - very high read peaks 60 GB/s



Instance Challenges

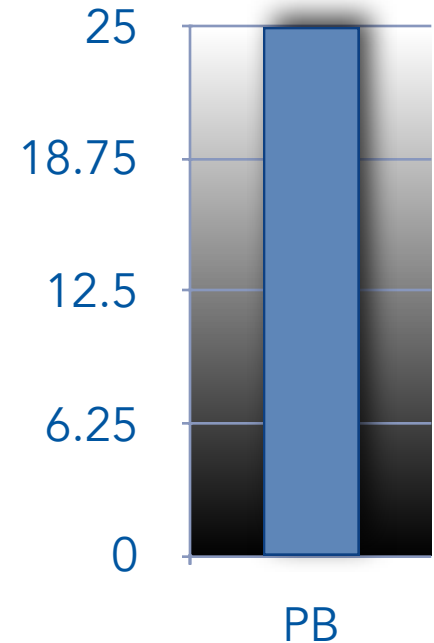
ATLAS

- complex usage
 - 1700 quota accountings
- (too) many directories - 30 M
- CDR input & export
- thousands of LAN connections
- full spectrum of use cases
- high (GSI) connection rates (100-200 Hz)
- many protocols in parallel



 Open Source Storage

 Open Source Storage



Instance Challenges

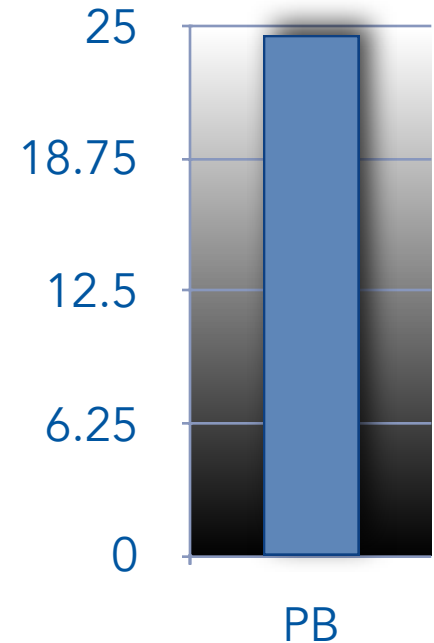


 Open Source Storage

 Open Source Storage

CMS

- complex usage
 - 1900 quota accountings
- moderate namespace size 60 M files
- CDR input & export
- thousands of LAN connections
- full spectrum of use cases
- high (GSI) connection rates (100-200 Hz)
- many protocols in parallel



Instance Challenges

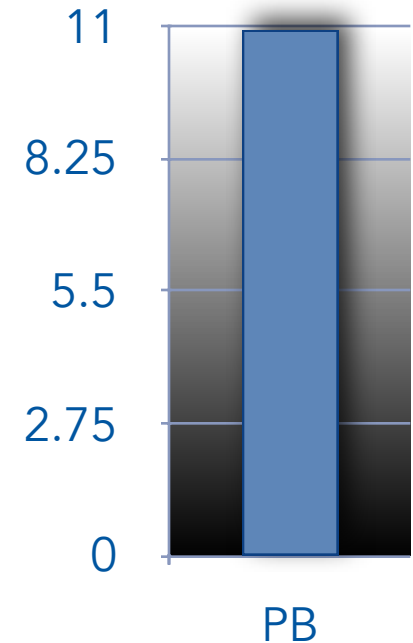


 Open Source Storage

 Open Source Storage

LHCB

- complex usage
 - 1600 quota accountings
- last LHC SRM customer
- smallest instance in terms of meta-data
- first instance running CITRINE in production & IPV6 enabled



Instance Challenges

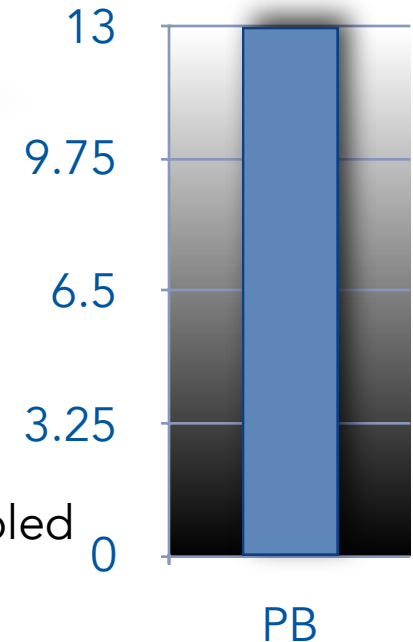


 Open Source Storage

 Open Source Storage

PUBLIC

- complex usage
 - 1400 quota accountings
 - many different experiments and admins
- large instance in terms of meta-data
- full spectrum of use cases
- second instance running CITRINE in production & IPV6 enabled



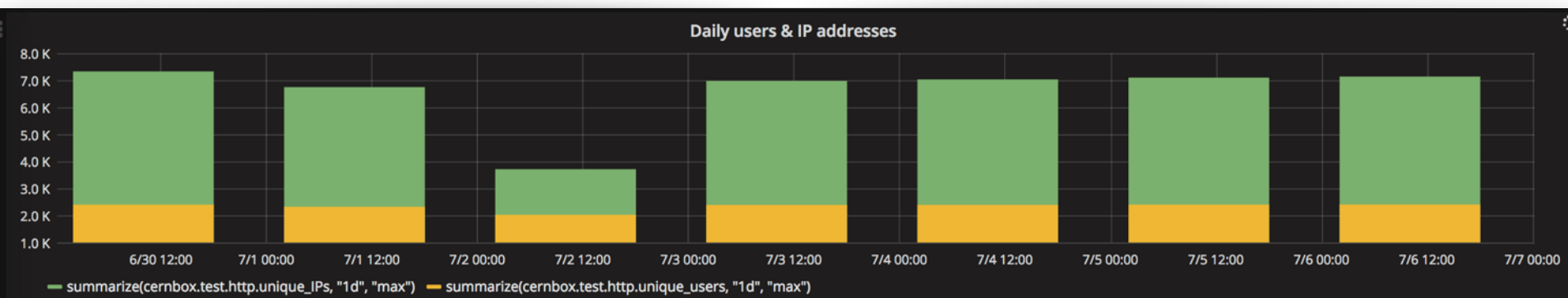
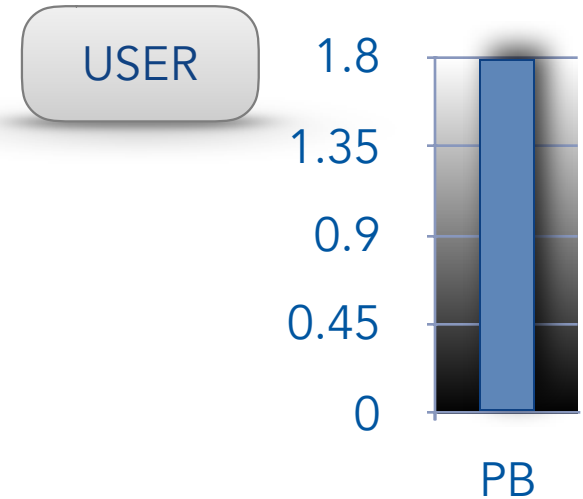
Instance Challenges



 Open Source Storage

 Open Source Storage

- complex usage
 - 10.000 quota accountings
- hundred millions of small files
- thousands of online users and devices
- full spectrum of (chaotic) use cases
- **CERNBOX** functionality
 - continuous background rate from sync clients

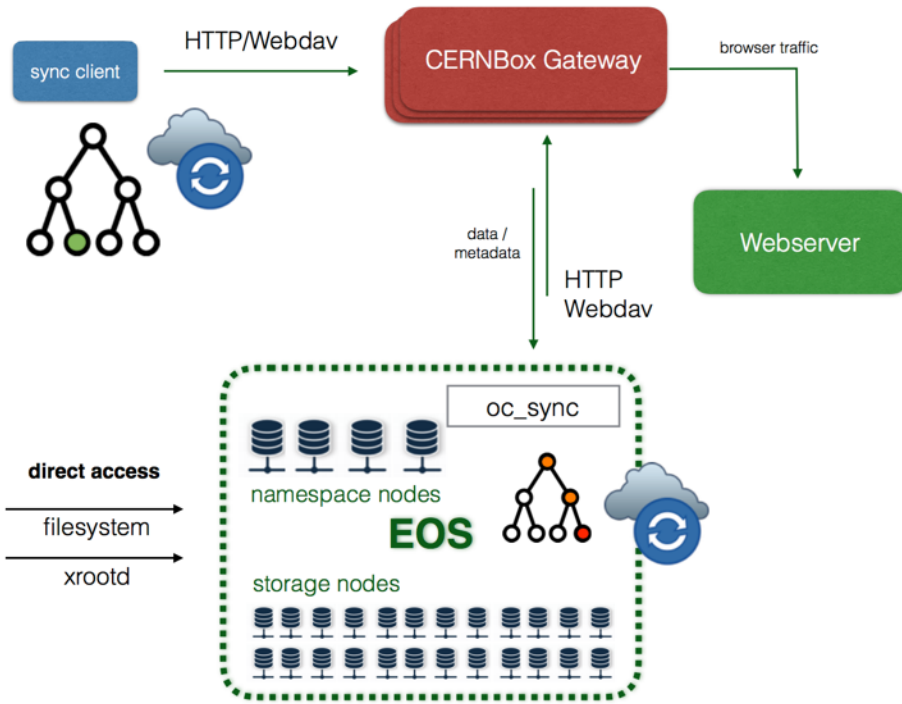




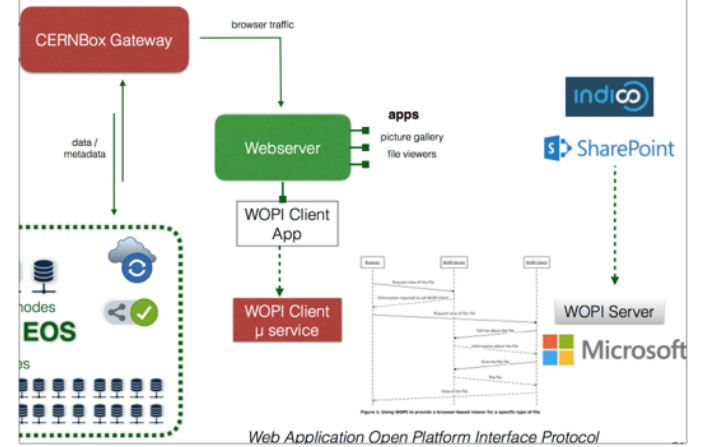
What is CERNBox

Sync & Share Platform = OpenSource Dropbox (OwnCloud)

Synchronization & Data Flow

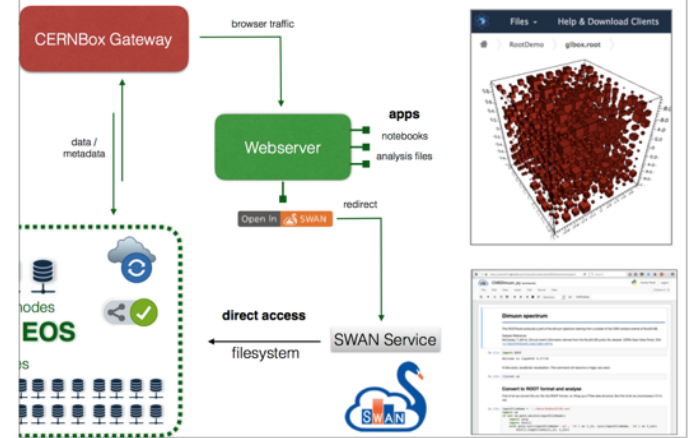


Web Apps, Office and Collaborative Editing



like GOOGLE docs

Web Apps and Scientific Computing



Jupyter Notebooks



CERN Disk Space Allocation

Experiment/Group	status Feb 2017	pledge 2017 (May)	request 2018
ALICE	17600	22400	27400
ATLAS	18500	25000	26000
CMS	22700	24600	26100
LHCb	7800	10900	12000

EOS [+ Castor]

2017 disk server commissioning done
48 x 6 = **288 TB** server **5 GB/s** disk IO
100 PB for ALL@CERN (divide by 2)



EOS community



21 external deployments of EOS

- many HEP sites
- non HEP communities
 - AARNET Australia
 - JRC Italy

Developer Team



Operations team: one service manager for all EOS, one service manager for CERNBOX



Developments

Development News

- faster service startup
 - namespace load time 2-6x
- master-slave failover & compaction issues resolved
- CITRINE release finally in production
- CI - continuous integration platform on gitlab
 - build pipelines
 - **RPM** builds on SLC6, EL7, Fedora, OS X (client)
 - **DOCKER** image build
 - **automated testing** on every commit
 - coming: **kubernetes** cluster setups with **long-term testing**
details can be found on the EOS workshop page

<https://indico.cern.ch/event/591485/>

Development News

- nicer table formatting

```
EOS Console [root://localhost] |/eos/pps/proc/recycle/> version
EOS_INSTANCE=eospps
EOS_SERVER_VERSION=4.1.25 EOS_SERVER_RELEASE=1
EOS_CLIENT_VERSION=4.1.25 EOS_CLIENT_RELEASE=1
EOS Console [root://localhost] |/eos/pps/proc/recycle/> space ls
```

type	name	groupsize	groupmod	N(fs)	N(fs-rw)	sum(usedbytes)	sum(capacity)	capacity
spaceview	default	50	50	1705	1700	518.88 T	4.47 P	4.4
spaceview	spare	0	0	121	1	1.67 T	435.82 T	3.9

- disk health monitoring S.M.A.R.T

type	hostport	geotag	status	status	txgw	heartbeatdelta	nofs	balan-shd	drain-shd	gw-queue
nodesview	p05798818t64278.cern.ch:1095	0513::R::0050::RA65	online	on	off		1 48	2	0	0

host	port	id	uuid	path	schedgroup	headroom	boot	configstatus	drain	active	scaninterval	health
p05798818t64278.cern.ch	1095	7577	542c2e23-bb5a-443d-8a85-b8eb6147e6fa	/data01	spare	25.00 G	booted	empty	drained	online	604800	Check
p05798818t64278.cern.ch	1095	7580	d863feca-44e8-49a3-bc1d-5374be4d14f2	/data02	default.1	25.00 G	booted	rw	nodrain	online	604800	Check
p05798818t64278.cern.ch	1095	7583	461d062-0a75-4511-0171-d4b01230e5f5	/data03	default.10	25.00 G	booted	rw	nodrain	online	604800	Check

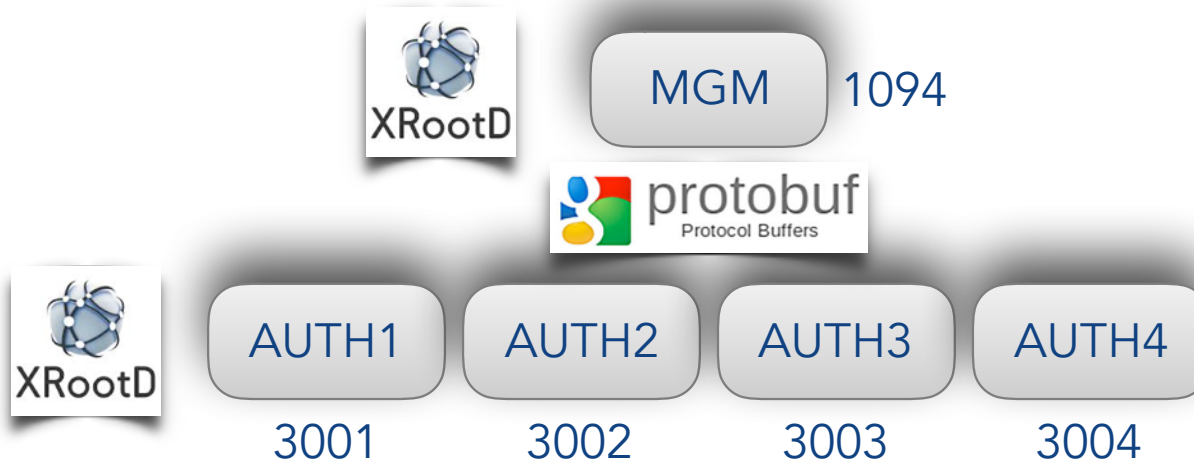
type	hostport	geotag	status	status	txgw	heartbeatdelta	nofs	balan-shd	drain-shd	gw-queue
nodesview	p05798818t49625.cern.ch:1095	0513::R::0050::RA65	online	on	off		1 48	1	0	0

host	port	id	uuid	path	schedgroup	headroom	boot	configstatus	drain	active	scaninterval	health
p05798818t49625.cern.ch	1095	7103	4714290b-1f1b-45e0-9f11-0eba5838b4c5	/data01	default.27	25.00 G	booted	rw	nodrain	online	604800	OK
p05798818t49625.cern.ch	1095	7110	bfef16ea-6113-4fe6-a2f9-9fa23de815c0	/data02	default.33	25.00 G	booted	rw	nodrain	online	604800	OK



Development News

- with Gerri Ganis boosted XRootD GSI plugin from 200 Hz to 1kHz handshakes
 - hit limit in ATLAS & CMS last month - results in very slow namespace responses
- deployed scale-out authentication service in CMS



```

AUTHPROXY_0 tcp -- anywhere anywhere statistic mode random probability 0.250000 /* 100 Authproxy probability routing */
AUTHPROXY_1 tcp -- anywhere anywhere statistic mode random probability 0.333333 /* 101 Authproxy probability routing */
AUTHPROXY_2 tcp -- anywhere anywhere statistic mode random probability 0.500000 /* 102 Authproxy probability routing */
AUTHPROXY_3 tcp -- anywhere anywhere statistic mode random probability 1.000000 /* 103 Authproxy probability routing */
    
```

Simple stateful load-balancing



1094

- CERN-IT extra-large disk server project
 - **8 x 24 x 6TB** disks connected to single front-end node [1.152 PB/node]
 - capacity/performance ratio ?
 - OS limitations handling 192 disks ?
 - RAID vs. ZRAID vs. Software EC
 - which network IF ?
 - which CPU type ?
 - TCO evaluation

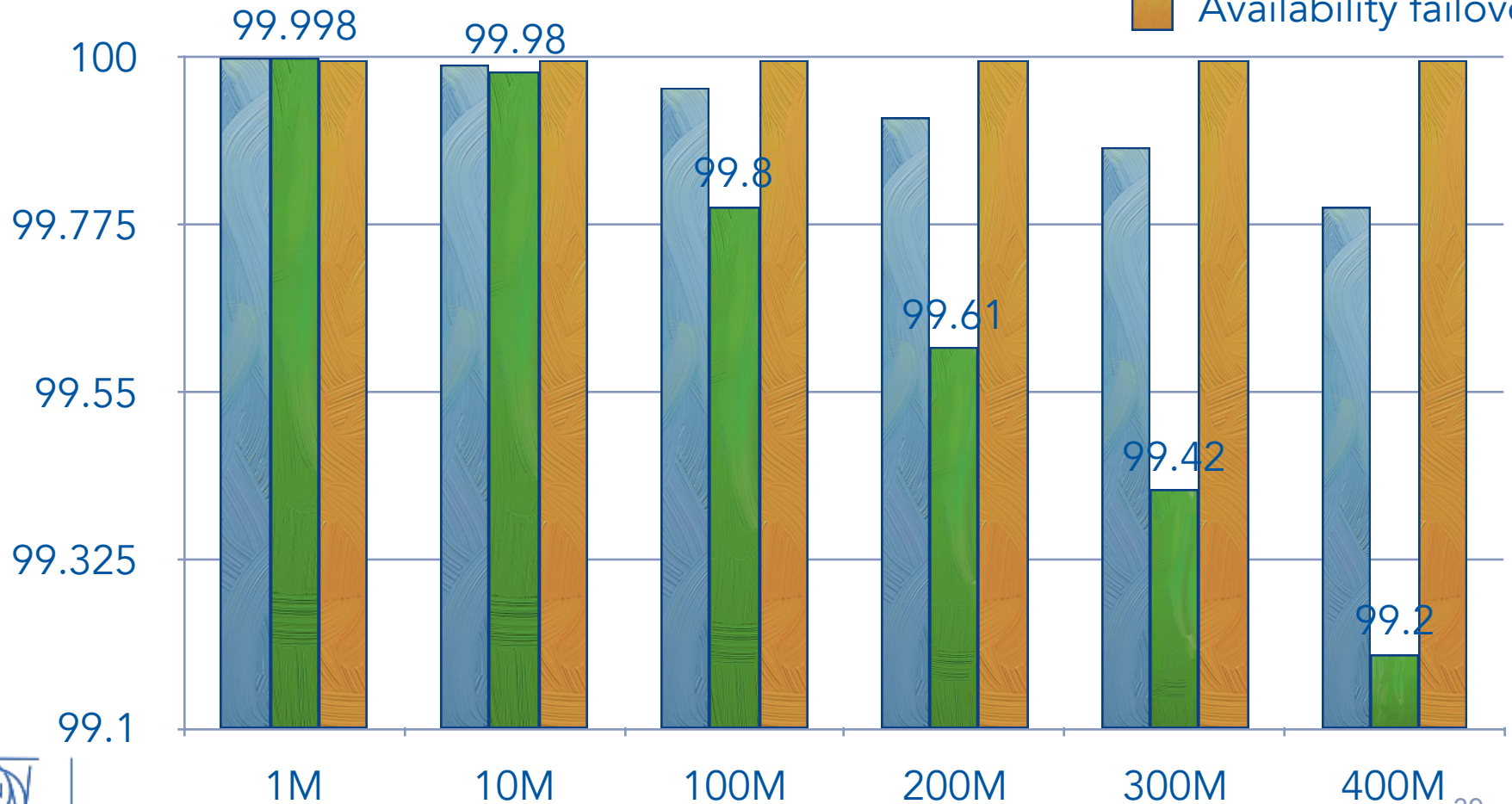


In-memory namespace



Availability [assuming uptime of 1 month]

- Availability defrag
- Availability frag
- Availability failover



New Parallel NS Boot



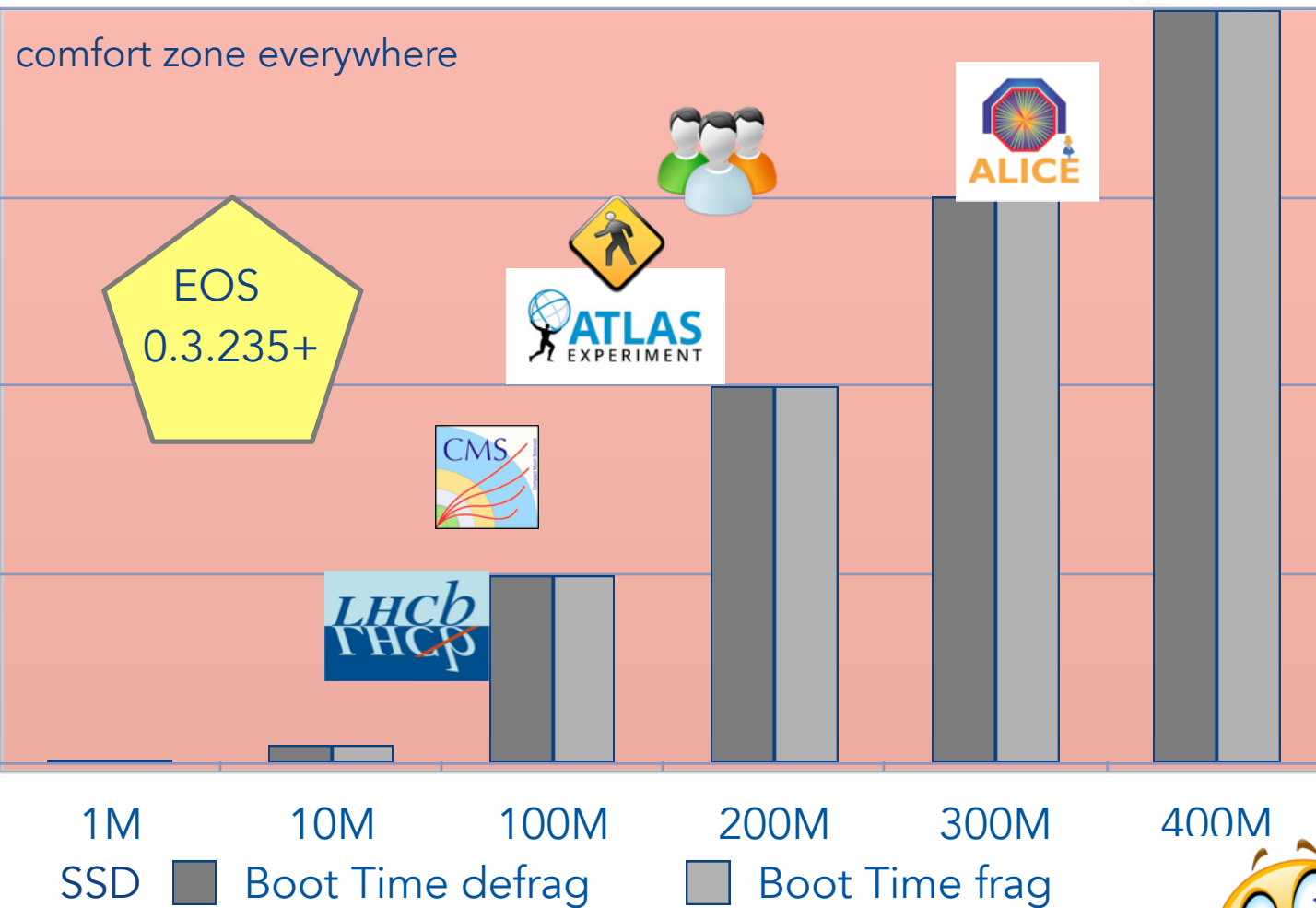
Open Source Storage

Open Source Storage



tolerable for user experience

Boot Time [s]



2x - 6x faster boot



EOS Architectural Evolution



Beryl Aquamarine
V 0.3.X



Citrine
V 4.X

read/
write

MGM
Master

MGM
Slave

read
only

META DATA

stateless



MGM

MGM

MGM

Persistency

FST

FST

FST

FST

DATA

FST

FST

FST

FST



KINETIC
Open Storage Project



ceph

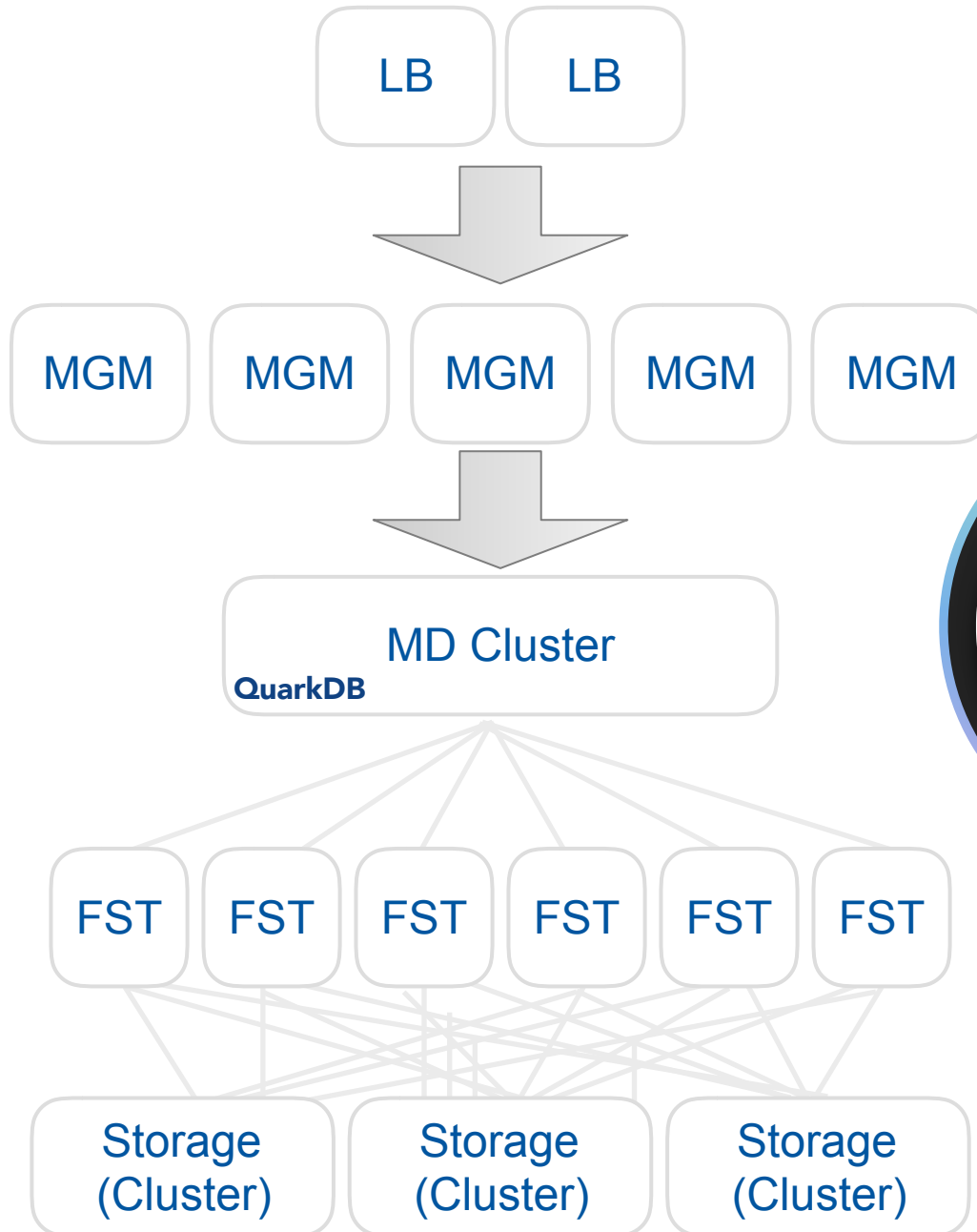


amazon.com




 Open Source Storage

 Open Source Storage



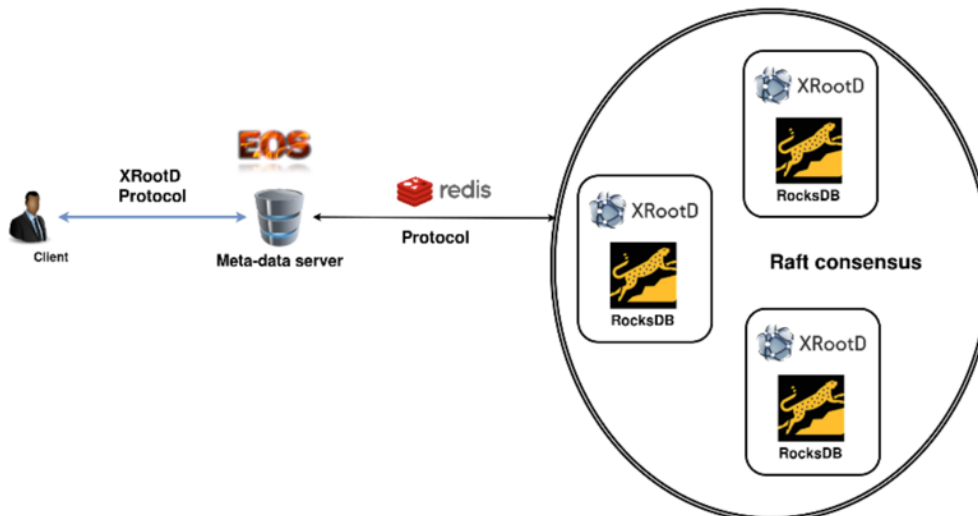
Meta Data Cluster for EOS

Namespace on top of a datastore

- Requirements: **consistent** low latency, scalable, very high rate of writes per second
- EOS to replace AFS at CERN, hold network home directories
- Needs to be reasonably performant for tasks such as
 - interactive usage
 - compiling
 - untarring archives the size of the linux kernel 

QuarkDB: a highly-available datastore

- Implement the minimum necessary to keep the system simple
 - QuarkDB runs as a plug-in to the XRootD server framework used by EOS
- A redis-like server on top of RocksDB
 - Support for a subset of the redis command-set: HASH, STRING, SET operations
- High availability through multiple **strongly-consistent** replicated nodes
 - Raft consensus algorithm to keep replicas in sync



- **10k** lines of C++ (including tons of tests)
- Preliminary benchmarks: peak of **100khz** 200-byte writes, **300khz** reads (non-replicated mode)
- Replicated performance currently **10–15 khz** writes – plans to improve through automatic sharding

2011

remote data store



Open Source Storage

Open Source Storage

Interface Evolution

remote data store

+

file transactional storage



2017

remote data store

+

file transactional storage

+

distributed filesystem behaviour



/eos



Information Exchange

Evolution

EOS has started 6 years ago as a remotely accessible data storage system with *posix-similar* interface. The interfaces has been extended to provide **file transaction functionality**. The most recent architectural change is to provide mounted filesystem semantics. This aims to help to take over more use cases of AFS to prepare for AFS phase-out. It is not clear yet that this will work with a native EOS storage system alone.



Generic HOME Directory Service



\$HOME Current Status

AFS



Ixplus
Ixbatch
linux managed

DFS



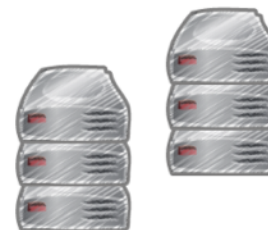
Terminal Servers
Windows managed

localdisk



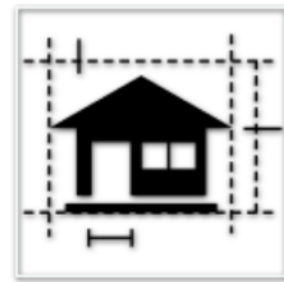
majority of users' devices

NFS



small private/experiment
clusters

\$HOME Future Vision



EOS\CERNBox



Terminal Servers
Windows managed

Ixplus
Ixbatch
linux managed



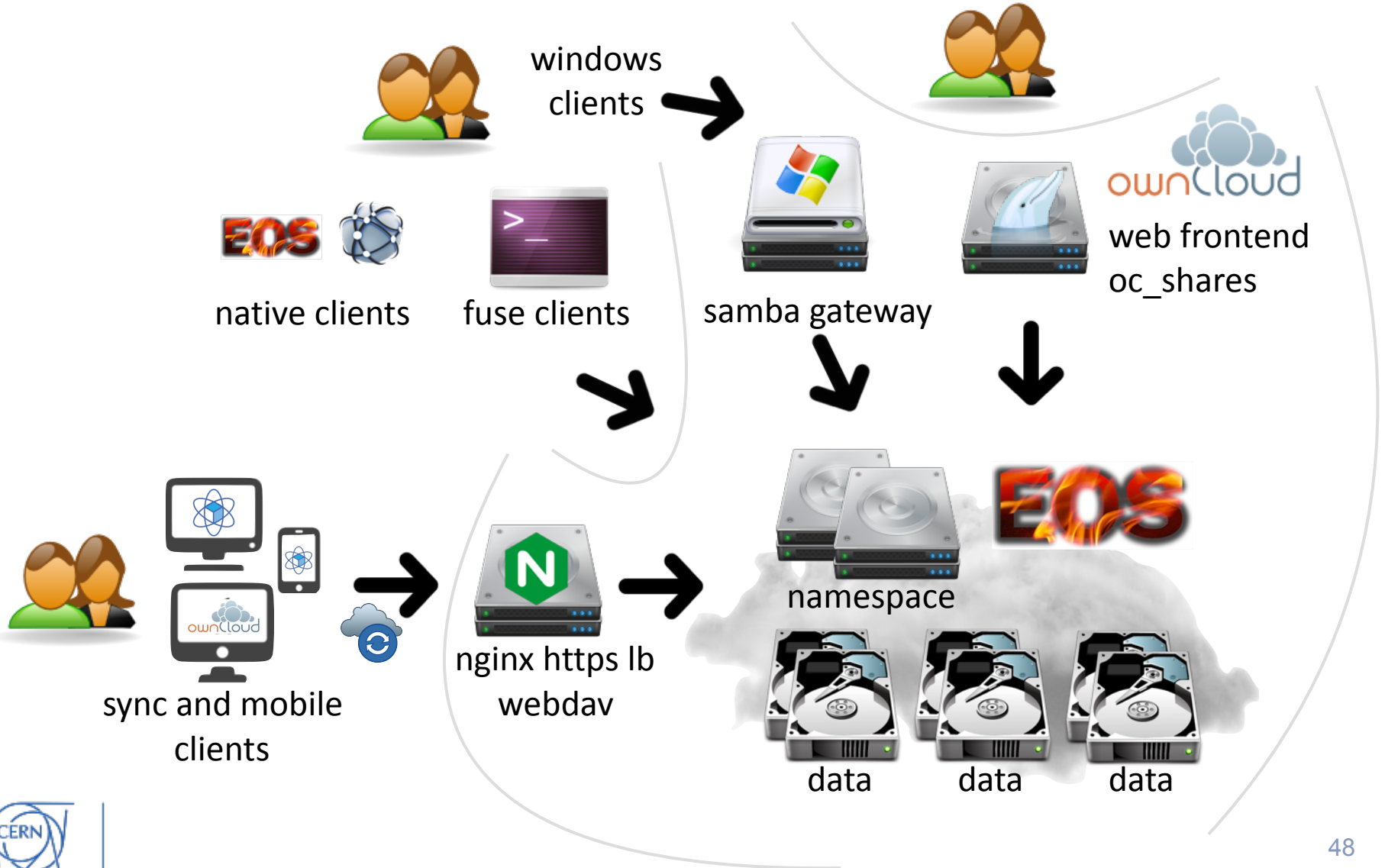
majority of users' devices

small private/experiment
clusters

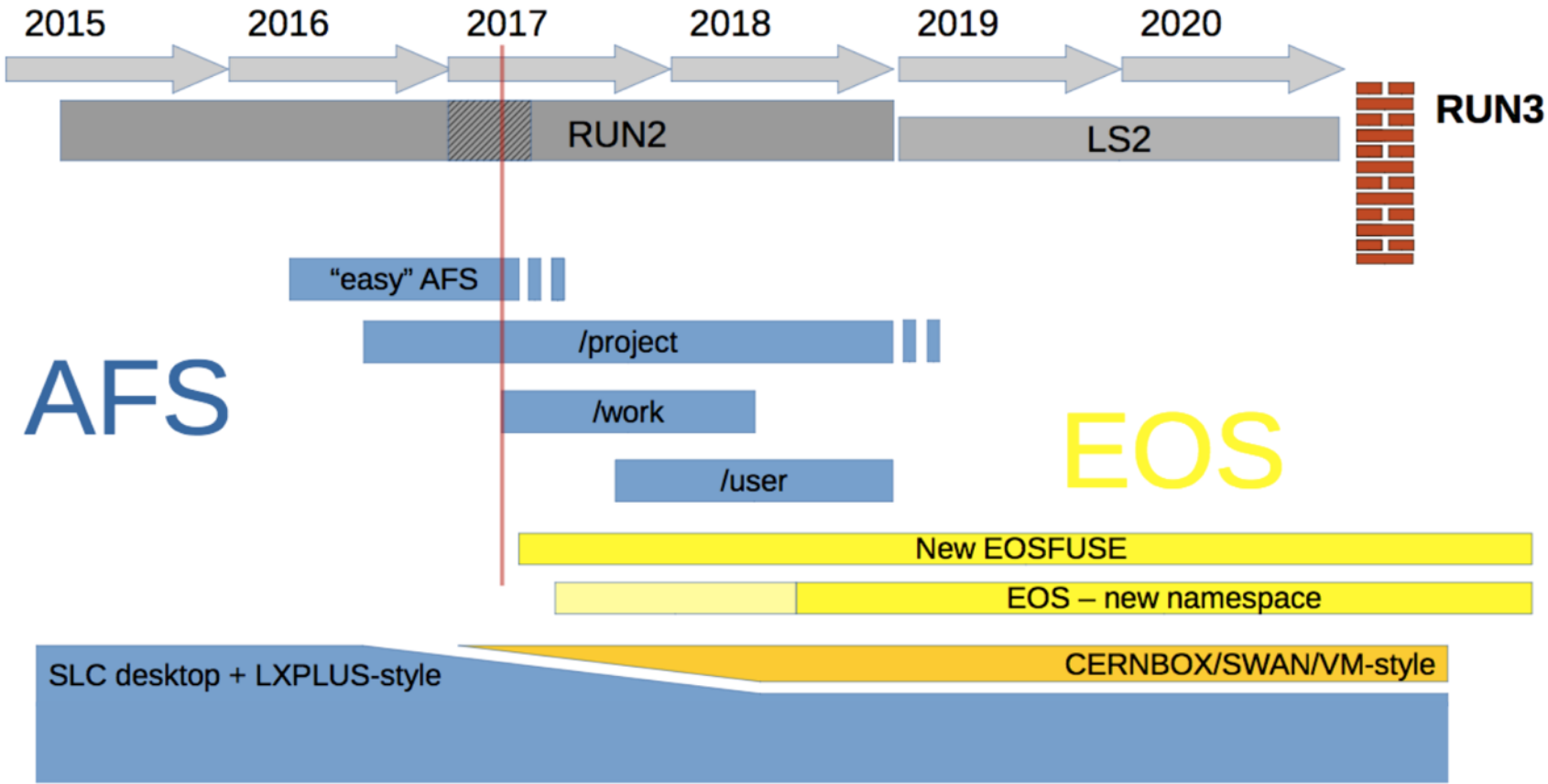




CERNBox Architecture



AFS Migration Plan

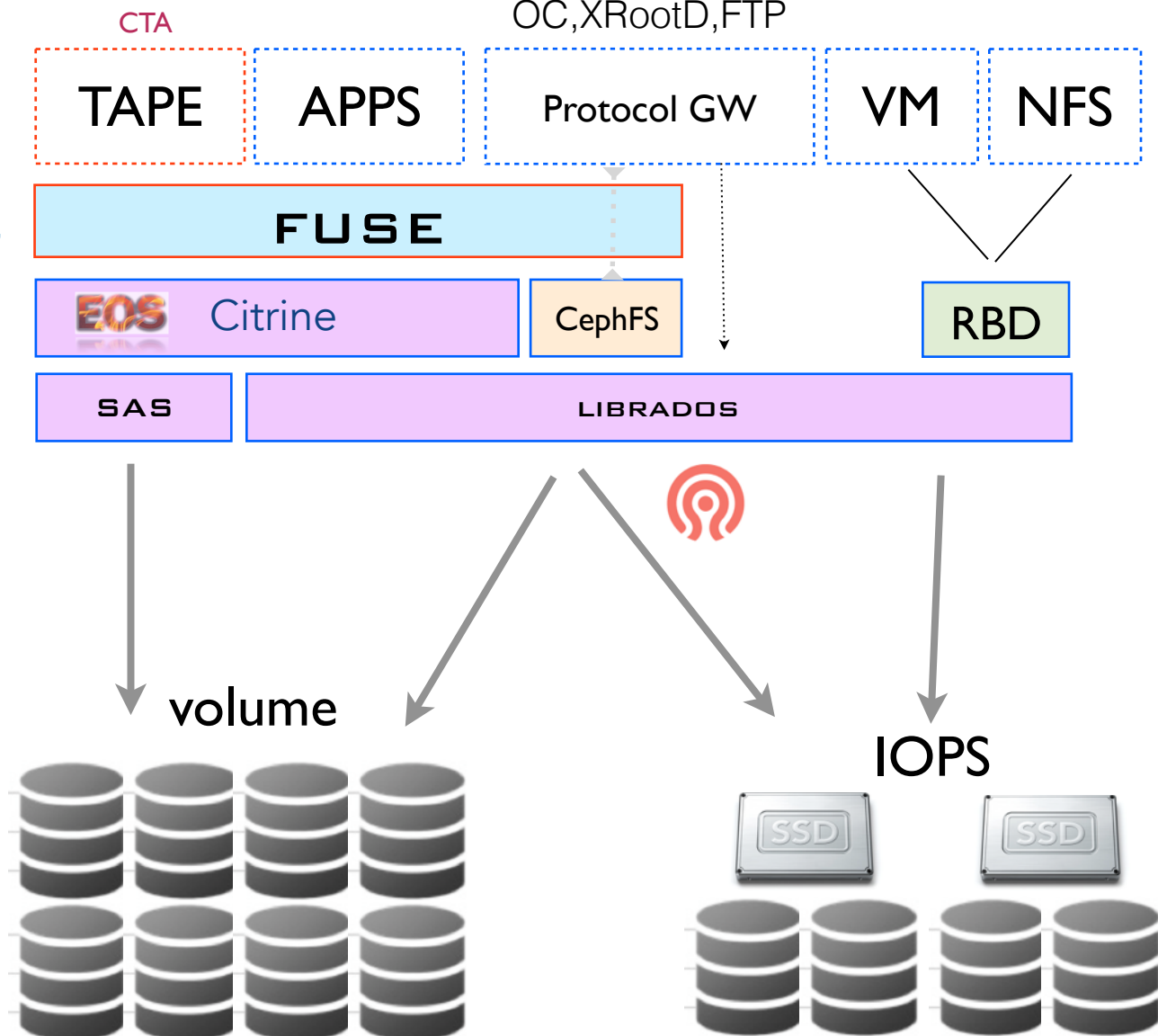




virtual groups/roles, acls, quota,
access/damage limitation, auditing,
krb5 + x509 auth

HTTP, DAV, S3,
OC, XRootD, FTP

Possible Plan B

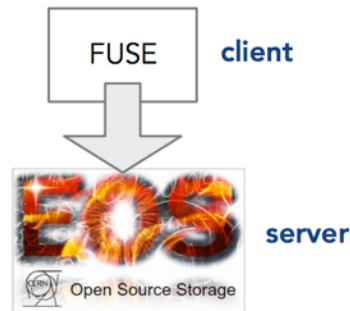


EOS FUSE Current Status

/eos mounted on lxplus and lxbatch nodes

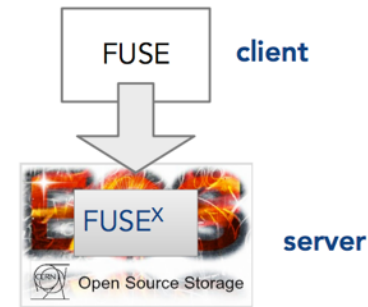
- significant amount of problems and obstacles
 - consistency, stability and kerberos integration
- experience triggered clean rewrite of FUSE client
 - implementation of a filesystem = challenging task !

V2 implementation



FUSE filesystem implemented as **pure client side** application without dedicated server side support.

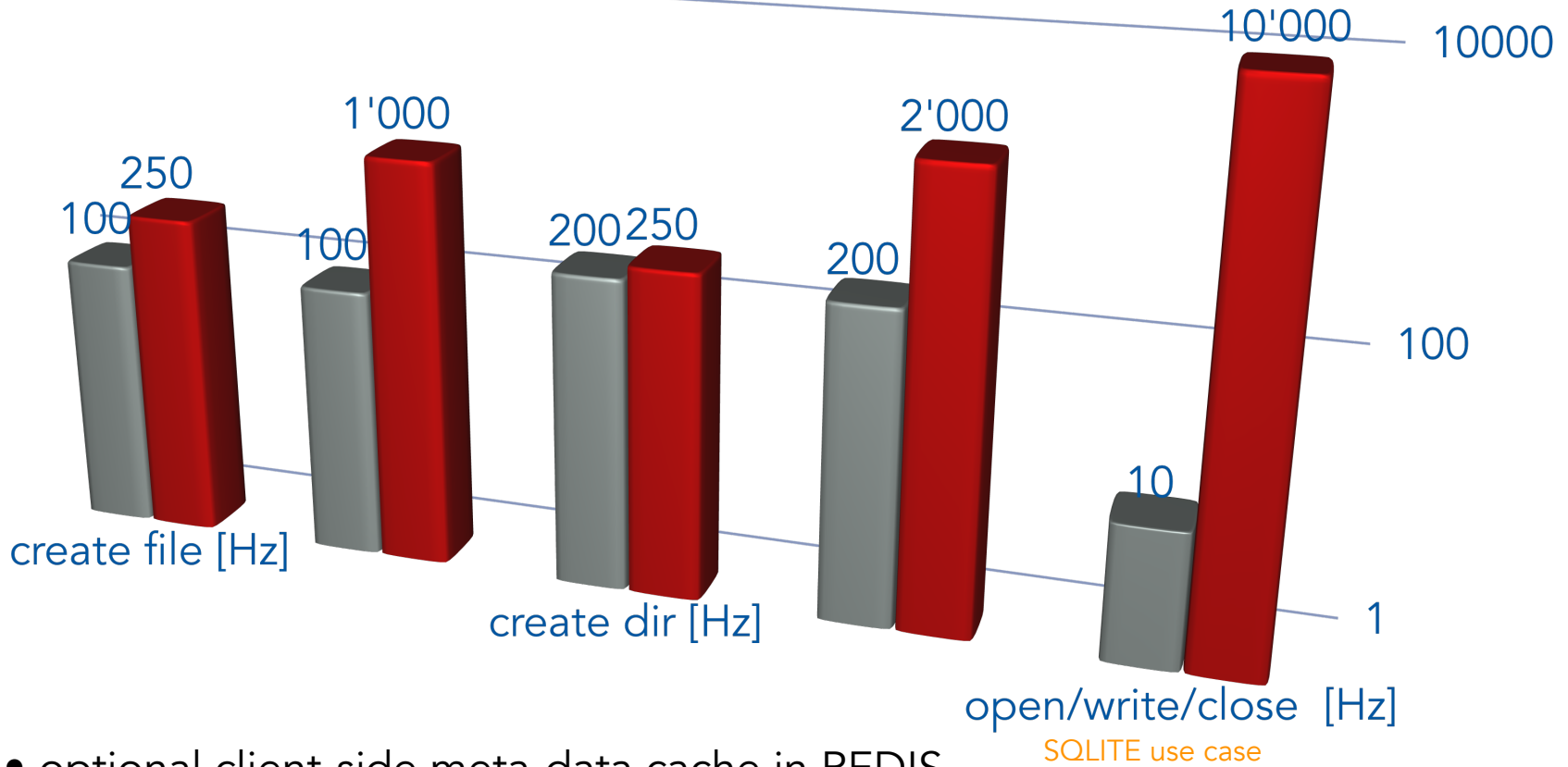
V3 implementation



Dedicated server-side support providing a fully asynchronous server->client communication, leases, locks, file inlining, local meta-data and data caching

First new FUSE performance impressions

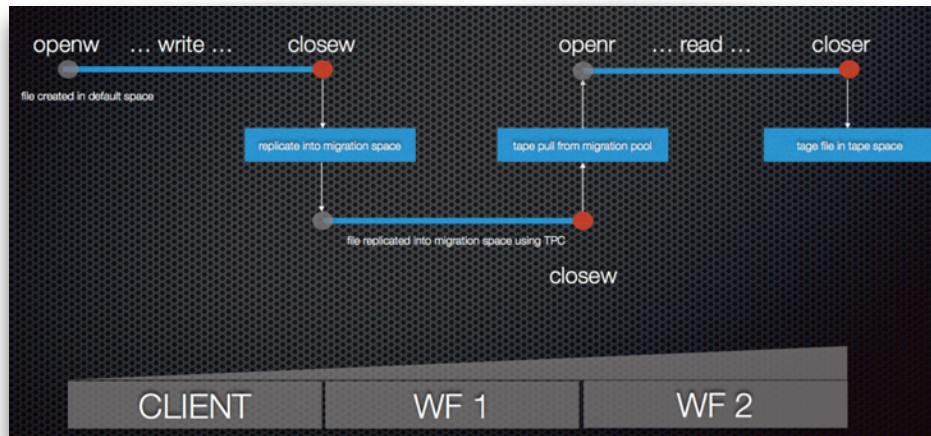
seen from one client



- optional client-side meta-data cache in REDIS
- configurable client-side data cache & file journal - leverage SSDs on batch nodes
- nfs4 exports via kernel nfsd

Storage Tiering in CITRINE

- generic support for file workflows in CITRINE



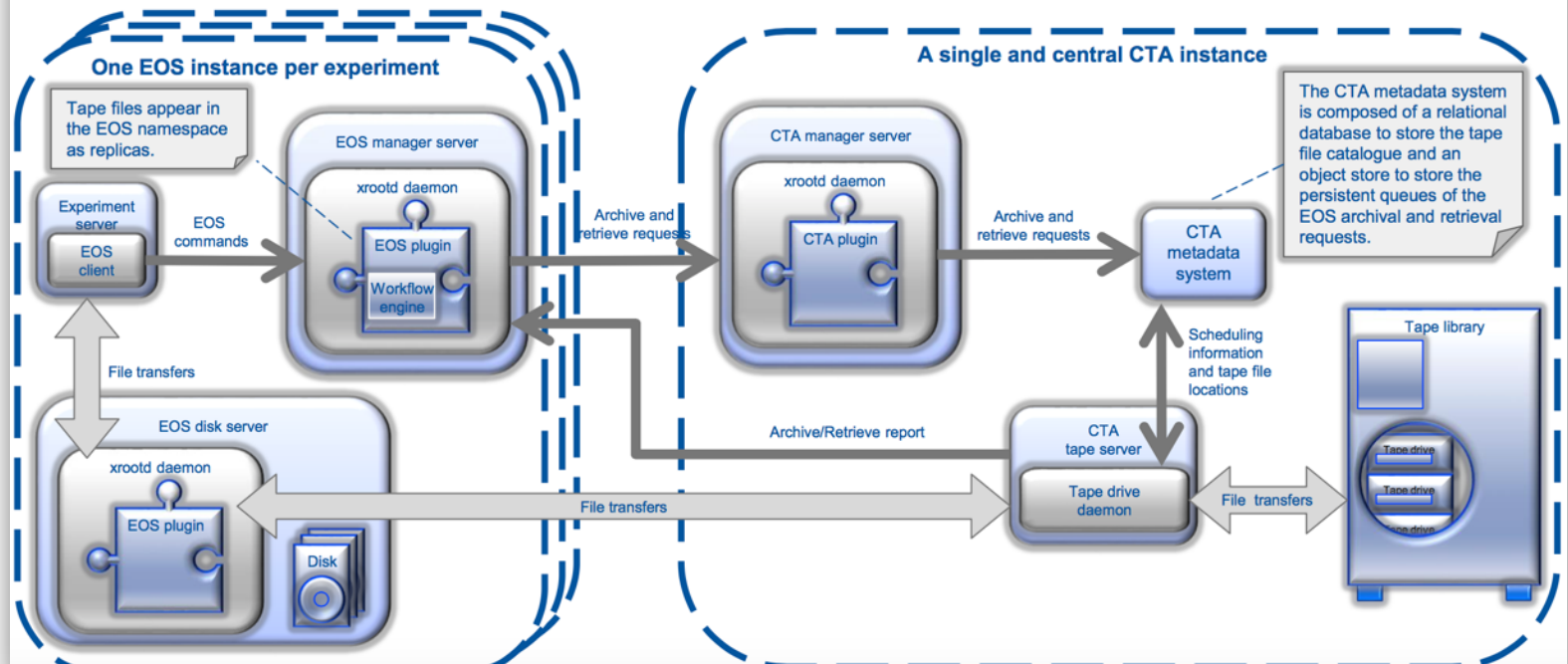
- connection to the **Cern Tape Archive** project as tape storage system
 - provides manual archiving or automatic migration/recall behaviour
 - CTA and EOS separate projects - exchangeable

EOS & CTA server communicate via protocol buffer bus



EOS+CTA architecture

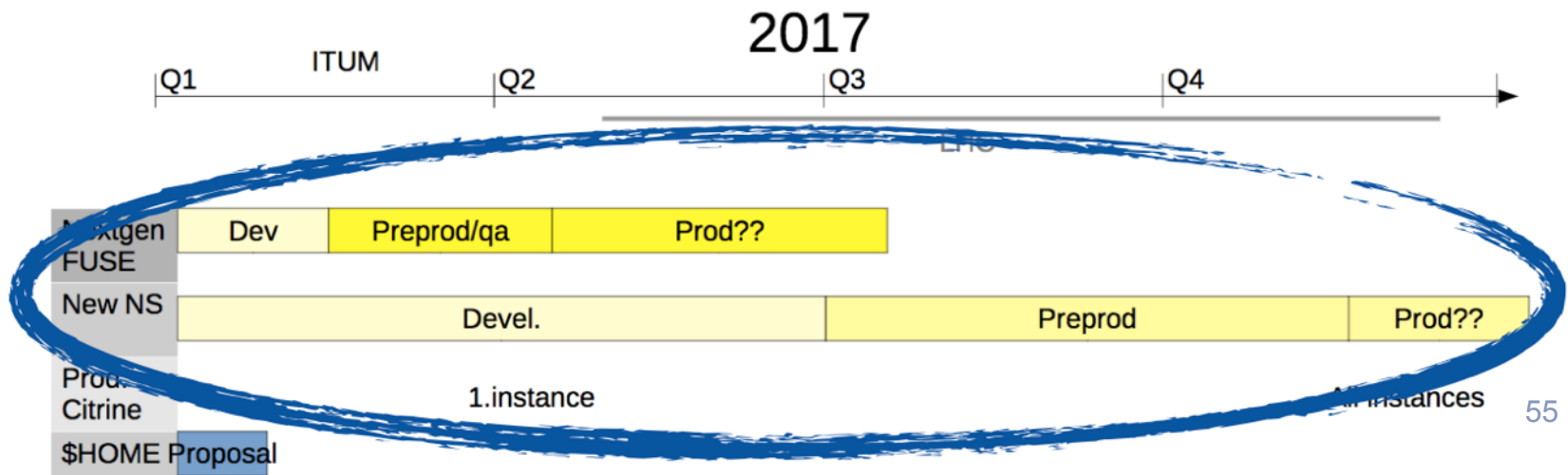
- CTA is integrated with EOS: all user interaction via EOS
- CTA tape files appear in the EOS namespace as file replicas
- CTA contains an internal flat catalogue of all tape files



Roadmap & Milestones



- new FUSE client deployment Q3
- CITRINE instances at CERN IPV6 /xroot4
 - 15th of May updated for LHCb still with in-memory namespace, PUBLIC instance followed in June
 - Q3-Q4 planning move from in-memory to QuarkDB namespace **namespace scalability - demonstrate multi billion ns**
 - add many 9's to availability 99.9 => 9999....
 - Cern Tape Archive milestone first system in 2018



Continuous Integration



<https://gitlab.cern.ch/dss/eos>

A screenshot of a web browser displaying the GitLab interface for the 'EOS' project. The browser's address bar shows the URL 'https://gitlab.cern.ch/dss/eos'. The page header includes the GitLab logo, the project name 'dss / eos', and a 'Sign in' button. Below the header is a navigation menu with options: 'Project', 'Repository', 'Registry', 'Issues', 'Merge Requests 0', 'Pipelines', 'Snippets', and 'Settings'. The 'Project' option is currently selected. Underneath, there are links for 'Home', 'Activity', and 'Cycle Analytics'. A dark grey notification banner at the top of the main content area reads: 'Changes in GitLab 9 affect API and CI: all users are invited to review OTG0037271'. The main content area features the 'EOS' logo, the project name 'EOS', and the description 'EOS project.'. At the bottom of this section, there is a 'Star' button with a count of '11', a dropdown menu for 'HTTPS', and a URL input field containing 'https://gitlab.cern.ch/dss/e'.


← ⓘ 🔒 | https://gitlab.cern.ch/dss/eos | 90% | 🔍 Suchen | ☆ 📁 📄 ⬇️ 🏠 ☰

☰ 🦊 dss / eos 🔍 [Sign in](#)

Project Repository Registry Issues Merge Requests 0 Pipelines Snippets Settings

[Home](#) Activity Cycle Analytics

🔔 Changes in GitLab 9 affect API and CI: all users are invited to review [OTG0037271](#)


EOS 🌐
EOS project.

★ Star 11 HTTPS https://gitlab.cern.ch/dss/e 📄

Continuous Integration

<https://gitlab.cern.ch/dss/eos>

Navigation: Project Repository Registry Issues Merge Requests 0 Pipelines Snippets Settings

Sub-navigation: Pipelines Jobs Schedules Environments Charts

Summary: All 643 Pending 0 Running 1 Finished 632 Branches Tags Run Pipeline CI Lint

Status	ID	Branch	Commit	Job Name	Progress	Duration	Time Ago
passed	#156738	master	85f53cb0	NS: Update the SyncTime and C...	! ✓ ✓ ✓	00:21:19	2 days ago
failed	#157339	master	a3126fda	MGM: changing event type for n...	! ✓ ✗ →	00:28:12	2 days ago



Continuous Integration



<https://gitlab.cern.ch/dss/eos>

A screenshot of the GitLab web interface for the EOS project. The browser address bar shows the URL 'https://gitlab.cern.ch/dss/eos'. The page header includes the GitLab logo, the project name 'dss / eos', and a 'Sign in' button. Below the header is a navigation menu with tabs for 'Project', 'Repository', 'Registry', 'Issues', 'Merge Requests', 'Pipelines', 'Snippets', and 'Settings'. The 'Registry' tab is highlighted with a red box and a blue circular callout that says 'CLICK HERE'. Below the navigation menu are links for 'Home', 'Activity', and 'Cycle Analytics'. A dark grey banner at the top of the main content area contains the message: 'Changes in GitLab 9 affect API and CI: all users are invited to review OTG0037271'. The main content area displays the EOS project profile, including the EOS logo, the name 'eos', and the description 'EOS project.'. At the bottom of the profile are buttons for 'Star' (with a count of 11), a dropdown menu for 'HTTPS', and the project URL 'https://gitlab.cern.ch/dss/e'.



Open Source Storage

Open Source Storage

Continuous Integration

Repository	Image ID	Size	Layers	Created
110028	82d4a1cd4	294 MB	17 layers	28 days
118648	c317f8b9a	311 MB	17 layers	8 days
111000	55894f2d0	313 MB	17 layers	7 days
109407	8b8631982	307 MB	17 layers	about 1 month
120982	c9942e6d3	309 MB	8 layers	2 days
98256	cdf654375	283 MB	17 layers	about 2 months
109407	b1490d872	280 MB	17 layers	29 days
121714	10af140b8	322 MB	8 layers	about 3 hours
120916	bcb015e7a	309 MB	8 layers	2 days



get a CITRINE image e.g.

docker pull gitlab-registry.cern.ch/dss/eos:121629

```
[root@eos-docker ~]# docker images
REPOSITORY              TAG                IMAGE ID           CREATED
gitlab-registry.cern.ch/dss/eos  121629           0b153e8b6506     5 hours ago
759.4 MB
```



EOS in DOCKER - 1 minute

git clone <https://gitlab.cern.ch/eos/eos-docker.git>

start a virtual instance with KDC, MGM, MQ, 6xFST

./eos-docker/scripts/start_services.sh

get a shell in the MGM service container

docker exec -it eos-mgm-test bash

run instance tests

eos-instance-test

```
[root@eos-docker tmp]# eos-docker/scripts/start_services.sh -i 0b153e8b6506
468d806d6f2d8ad9398006818d0df281ac73621ece9b9c739c36596c2a05fb17
700257d7437b3a0b72f0ecdccfae95557932f5b73871ad27b1ee5ccd7dede03a
Starting kdc... Done.
Initing kdc... Done.
Populating kdc... added admin1@TEST.EOS with password "tDp5mfkjBx"
added host/eos-mgm-test.eoscluster.cern.ch@TEST.EOS with password ">CzjpraSdm"
Done.
378267ed609c0806a73bc9a4163aff27573ed089cf29d84eb81043d5876c635d
2ecea01ccabb84442a024ef458dd0117519cece3b77b62085fala2a71b65bd42
success: set vid [ eos.rgid=0 eos.ruid=0 mgm.cmd=vid mgm.subcmd=set mgm.vid.auth
m.vid.cmd=map mgm.vid.gid=0 mgm.vid.key=<key> mgm.vid.pattern=<pwd> mgm.vid.uid=0
2d1ba4c7111c933d89d6ad282643cb4287b11c47f6a7a94149cbbba1763fae651
02398a7e926966300016004397e59d419c26a13c256f7a64e13969ee222f7512
4e906031e8eb8b61bea2692aa8aec41c9921136676daed4eccd7aa6ebef51f21
7f42121461ed5c28961e87689fdcc0671f72d19a7936473387a3dc112e4e3607
75e76e25e2ac435cffde5476779162775cbf705720b10f274dc7e378e39a8446
152bffdaafdac36af00a69b3dd3b7e68eacbf9b54f479a15f3e09c6d7dcb915
Starting fst1 ...
Starting fst2 ...
Starting fst3 ...
Starting fst4 ...
```



Federations

WEST AUSTRALIANS
Complete the Union
by voting
YES!!!

FEDERATED AUSTRALIA
ONE PEOPLE
ONE FLAG
ONE DESTINY
377,000 AUSTRALIANS!!
have voted
"YES"

NOTE THIS
Every adult, resident 12 months in W.A. is entitled to Vote
If not on 1899 Roll a Voters Certificate must be obtained
All Ladies must obtain Certificates on or before 28th July 1900
Personal application is absolutely necessary.
See Advertisements in Daily Papers for Registrars Offices
*Obtain your Certificate and Vote **YES!!!***

All the Federated Canadian States now Prosperous.
Newfoundland stood out now Bankrupt.

Existing Customs collected in Colonies already Federated (average) per head £1:16:4.
Collected in Western Australia per head £5:0:5.

Route of Mail Steamers
FREMANTLE First and Last Port of Call.
PERTH
ADELAIDE
MELBOURNE
SYDNEY
BRISBANE
TRANS-CONTINENTAL RAILWAY

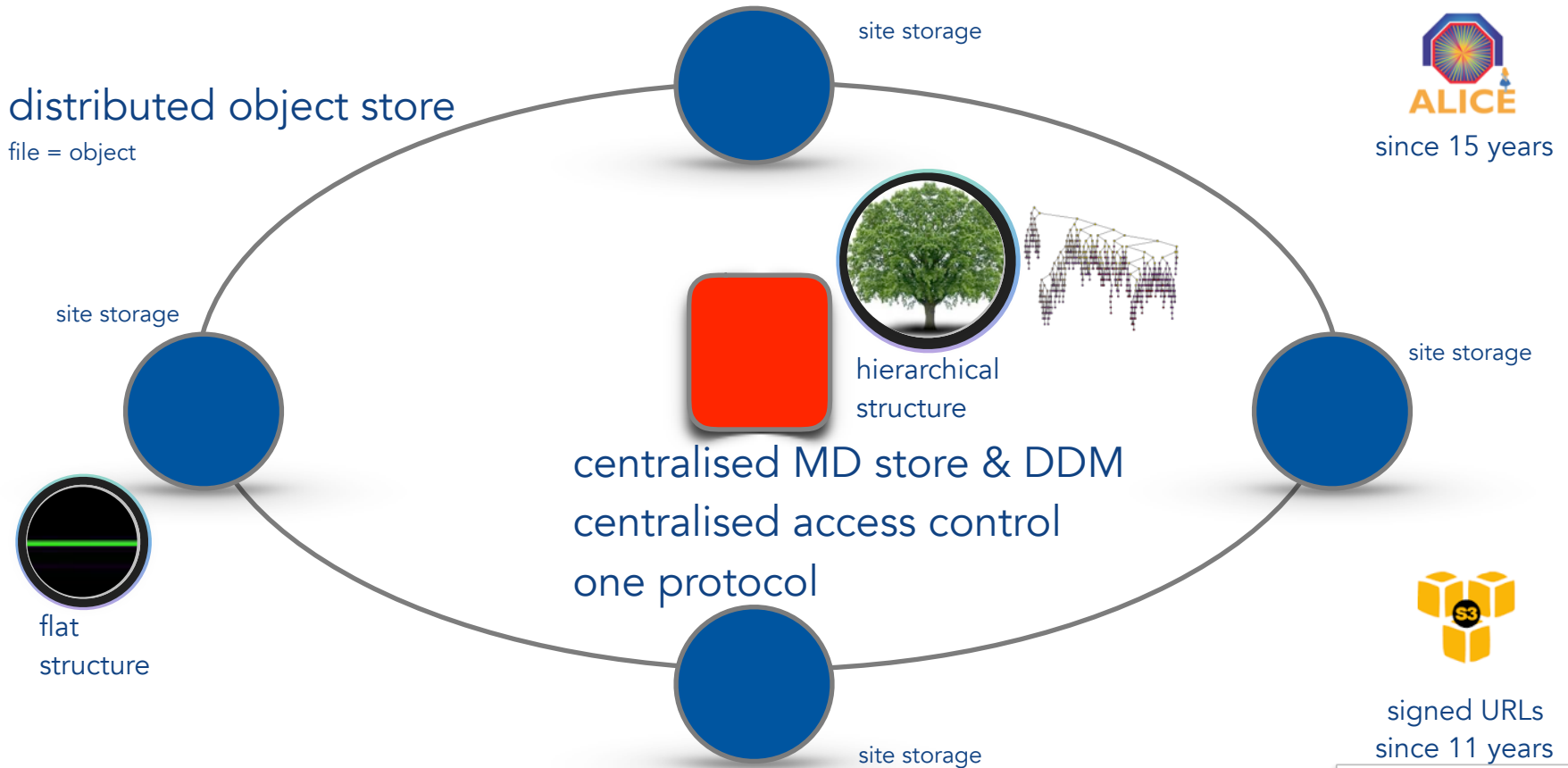
noted "YES"

Storage Federations



- driving idea is to
 - **reduce** the number of storage services to manage
 - **bundle** many small resources into a bigger single resource
 - **reduce** operational effort
 - few complex services (namespaces)
 - many trivial services (*object* storage servers)

Global Federation / Global Storage System



every access contacts central service

Simple.

Hybrid Federation



distributed file store

file = object + acl

decentralised
access control
many protocols

centralised DDM

site storage

site storage

site storage

site storage

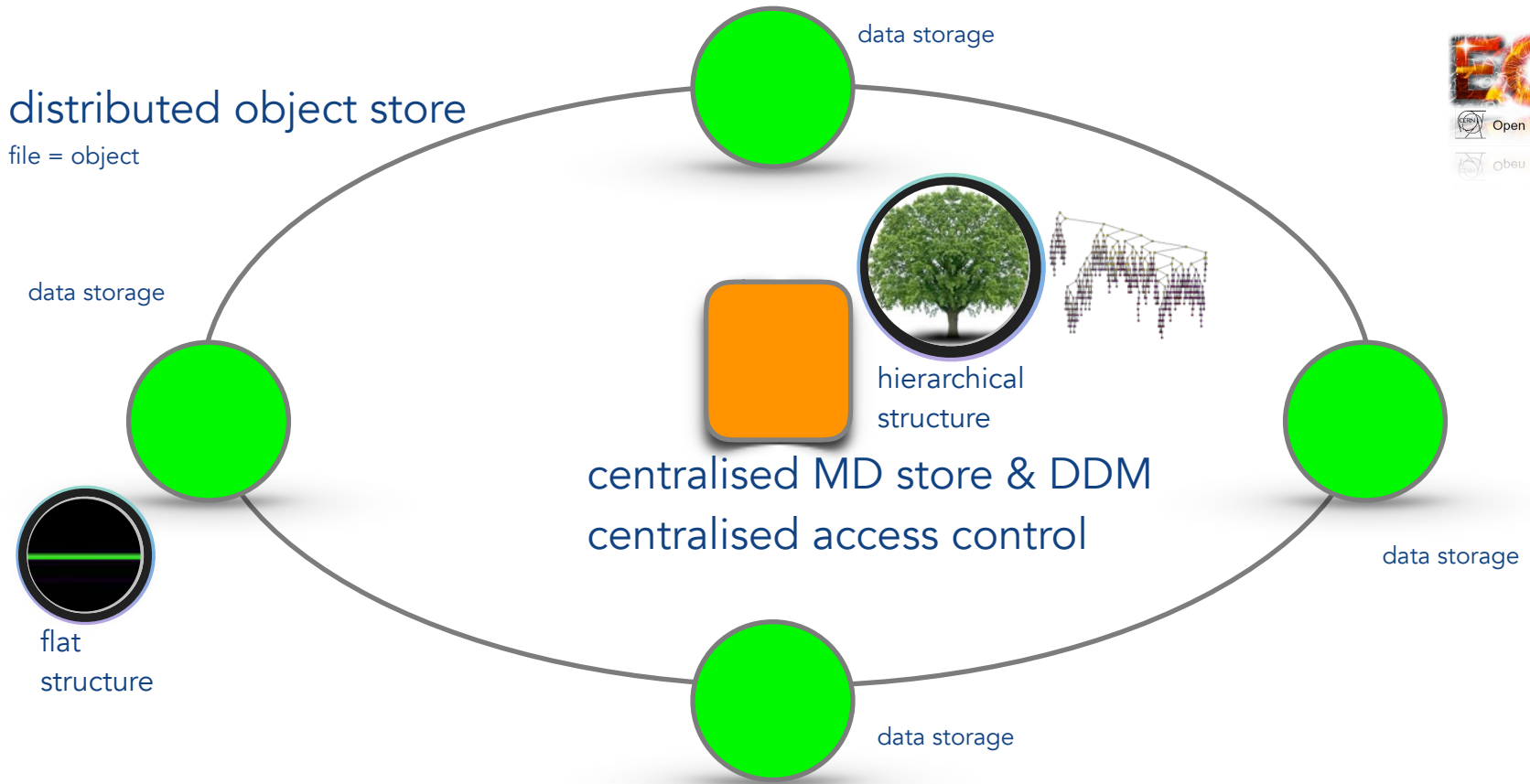


COM
PLEX

data distribution and discovery is centralised
access control is local



Storage-Level Federation



every access contacts storage entry point
batch problem: locality is hidden by the storage system

EOS Storage Federation

3 types of sites

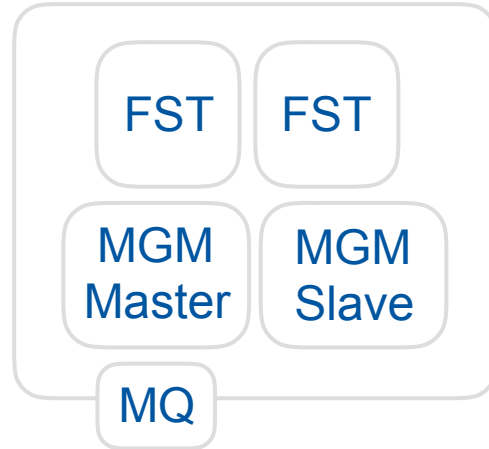


Open Source Storage



Open Source Storage

storage site **type 1**



storage server

active/passive meta data server

MGM
Follower

FST

FST

FST

storage server

storage site **type 2**

FST

FST

storage server

storage site **type 3**



EOS Placement Strategies

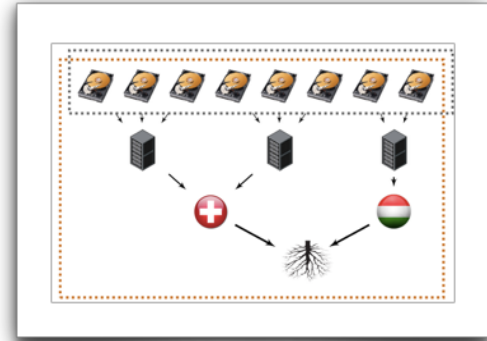


Open Source Storage



Open Source Storage

- CITRINE release provides **geographic-aware scheduling**
 - placement and access policies configurable for each directory
 - client and servers are geo-tagged (client:subnet server:configuration value)
 - access files as 'close' as possible
 - placement policies e.g.
 - hard-coded replication to defined locations
 - two replicas close
 - one close, one randomly
 - two replicas at maximum distance
 - ...



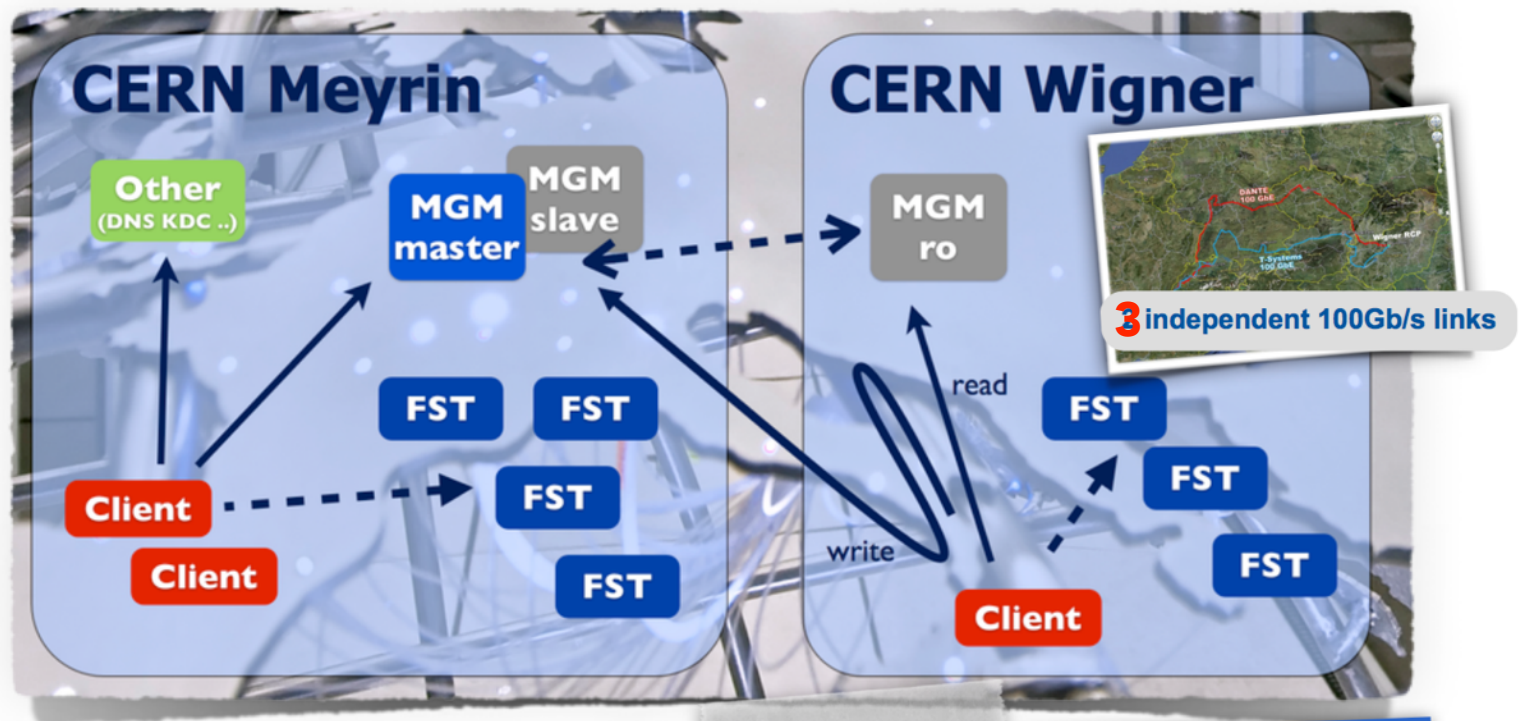
EOS Federations

Examples

EOSALICE: 159 storage nodes at CERN & 85 at WIGNER

placement policy: if possible maximum distance e.g. one replica at CERN, one at Wigner

Wigner Computer Centre



```
[root@lxbse15c06 ~]# ping p05153065491511
PING p05153065491511.cern.ch (188.185.224.50) 56(84) bytes of data.
64 bytes from p05153065491511.cern.ch (188.185.224.50): icmp_seq=1 ttl=58 time=22.0 ms
64 bytes from p05153065491511.cern.ch (188.185.224.50): icmp_seq=2 ttl=58 time=22.1 ms
64 bytes from p05153065491511.cern.ch (188.185.224.50): icmp_seq=3 ttl=58 time=22.1 ms
64 bytes from p05153065491511.cern.ch (188.185.224.50): icmp_seq=4 ttl=58 time=22.1 ms
```

22ms rtt



EOS Federations Examples

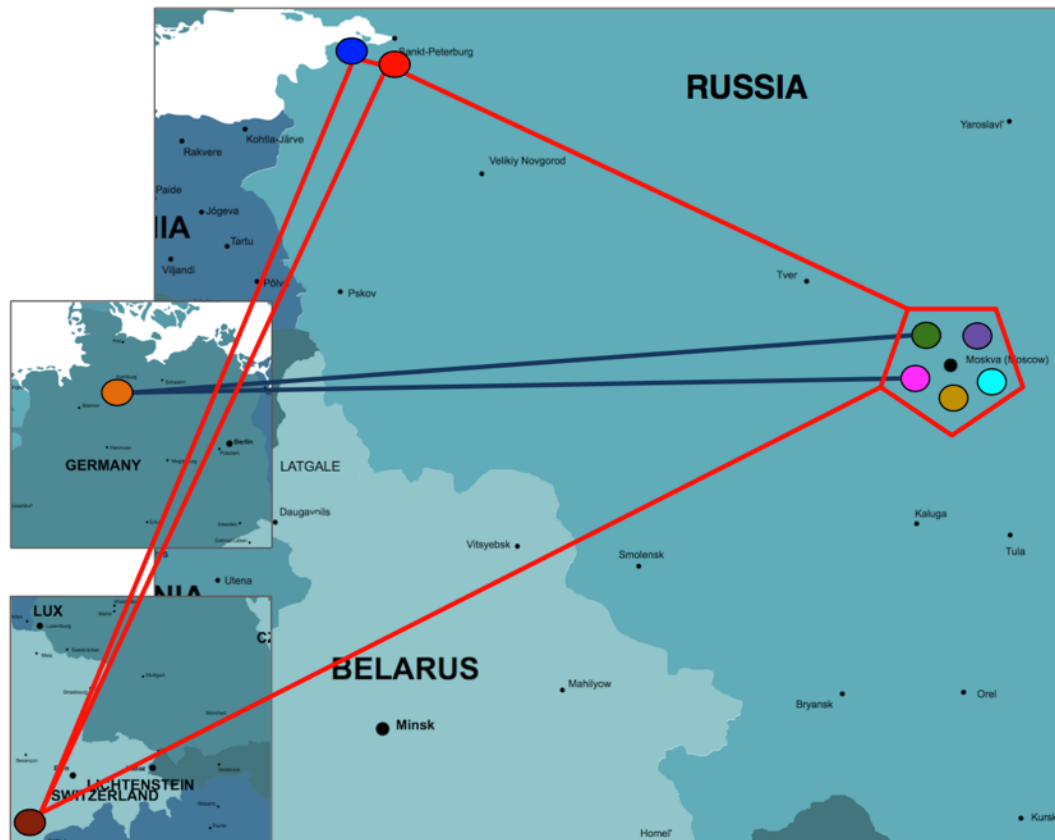


National Research Centre (NRC)
"Kurchatov Institute"



Big Data Technologies Laboratory
<http://bigdatalab.nrcki.ru/>

Federation topology



Andrey Kiryanov



EOS Federations Examples



AARNET Federation in Australia with 65ms latency

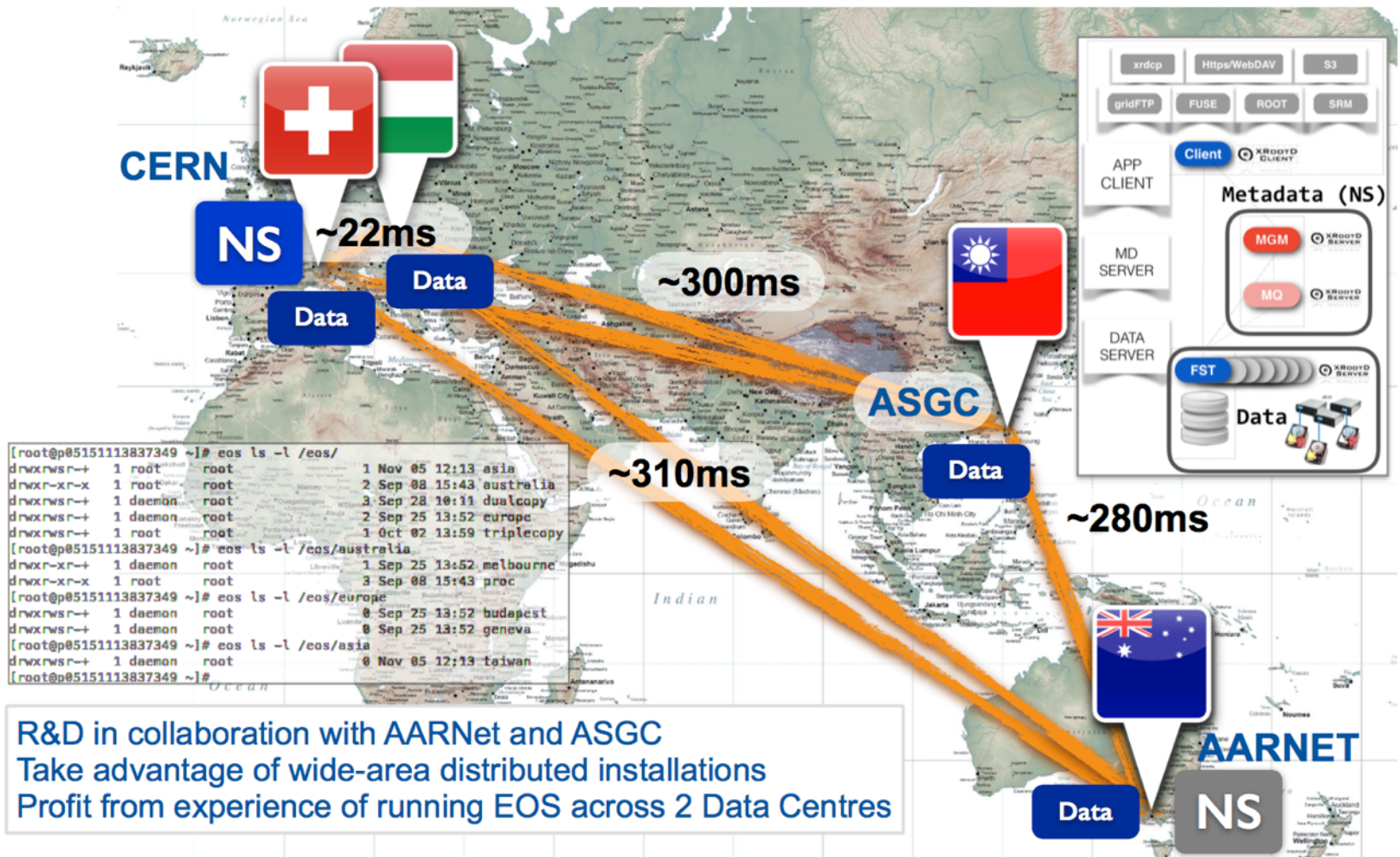
- three storage locations
- one RW/RO namespace pair
- one remote RO namespace



EOS Federations

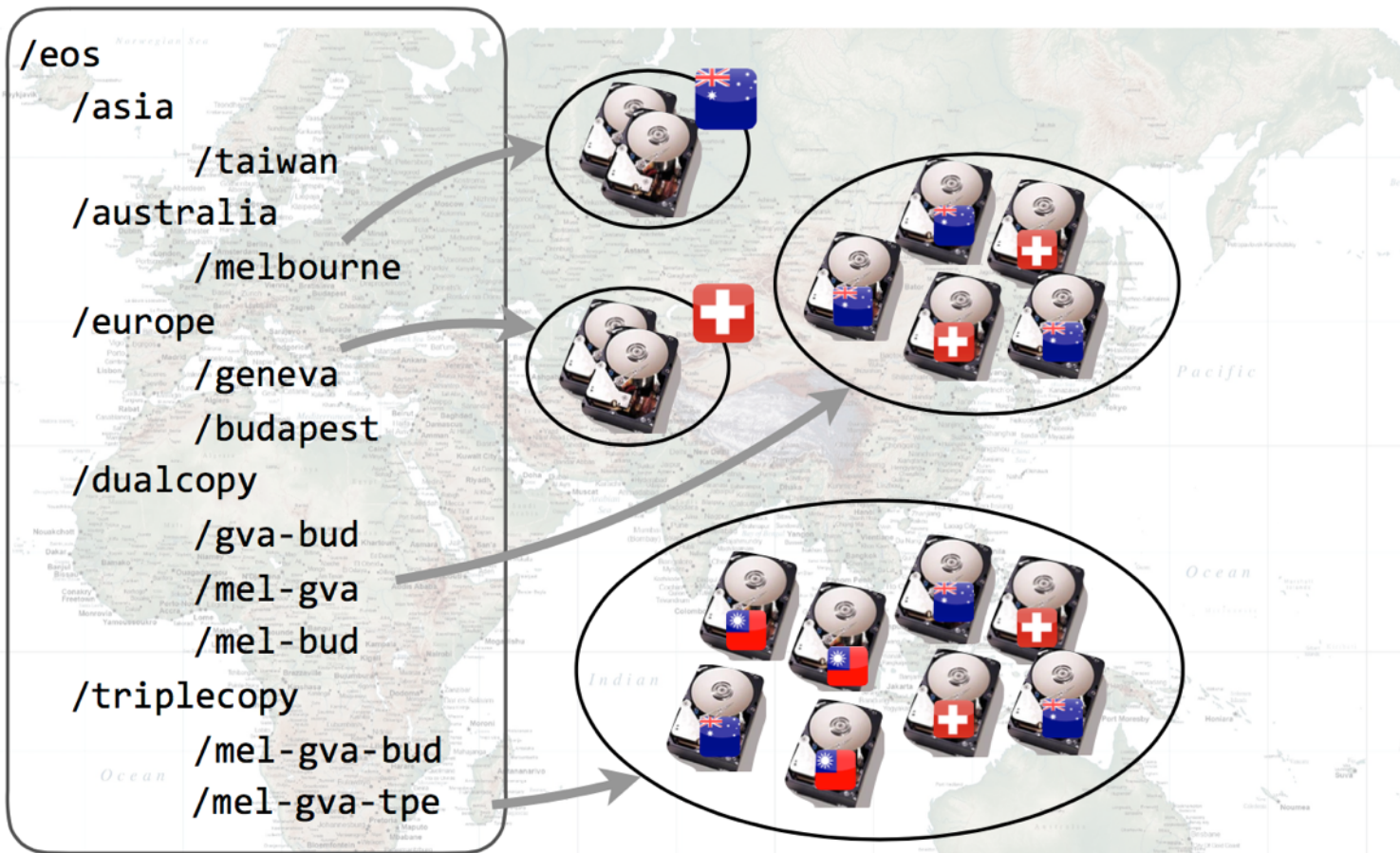
Examples

Worldwide distributed storage system with extreme latencies - R&D prototype



EOS Federations Examples

Worldwide distributed storage system with extreme latencies - R&D prototype



Storage pools were created with filesystems from all four sites. Files were replicated according to the different configured policy (e.g. 3 replicas: MEL-GVA-TPE).

EOS Federations

when to use them

- **doable** if all storage servers have sufficient bandwidth between each other and batch resources in the federation

“the WIGNER experience”

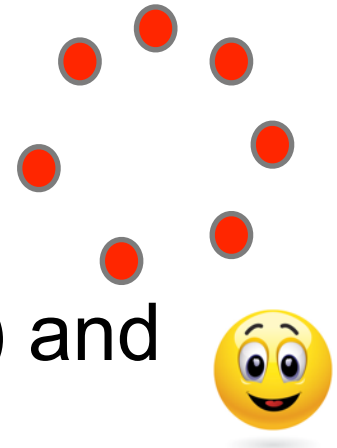


- two site federated setup complicated operations
 - implies 100% storage overhead [2 replica]
 - cannot use erasure encoding for recent data
 - network links become saturated
 - cannot guarantee 50:50 resources CERN:Wigner
 - TCP window scaling, a lot of network tuning and network problems to debug
 - job efficiency worse for 22ms latency than in LAN (depends on application)

EOS Federations

when to use them

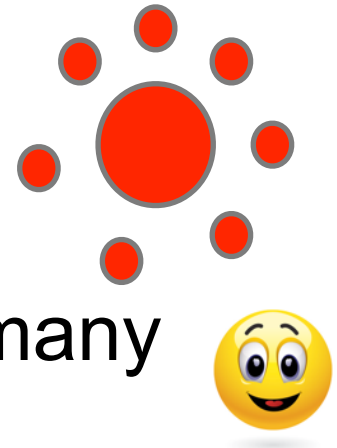
- **local cloud federation model**
(regional federation)
 - storage peers are close ($O(2)$ ms) and have sufficient network bandwidth
 - model requires efficient remote data access because:
 - impossible/inefficient to have a replica at each geo-graphic location
 - remote access will occur frequently



EOS Federations

when to use them

- **satellite federation model**
(T1/T2 federation)
 - a large resource is federated with many significantly smaller resources
 - main storage volume and CPU provided by single resource
 - T1 provides namespace and storage
 - T2 attach to T1 namespace



Storage Federations

when not to use them

- federations of several big resources are orthogonal to the GRID processing model
 - job scheduling should be location aware
 - a federation abstracts locality and network becomes a bottleneck between large resources
- storage federation have to be complemented and integrated with CPU resource federations - no free lunch - sounds nice to have only few big virtual sites but ..



my personal conclusion:

- the global federation model is the most space efficient federation (global policy for replication) - not operation wise
- regional federations of small contributors to storage and CPU make absolutely sense and help to reduce operational effort
- large resource federation have too high impact on volume and job efficiencies





for HL-LHC



 Open Source Storage

 Open Source Storage

transformation →

Exabyte-scale
Object
Storage



CERN OpenSource core functionality



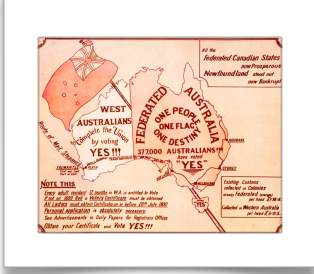
non-CERN OpenSource core functionality



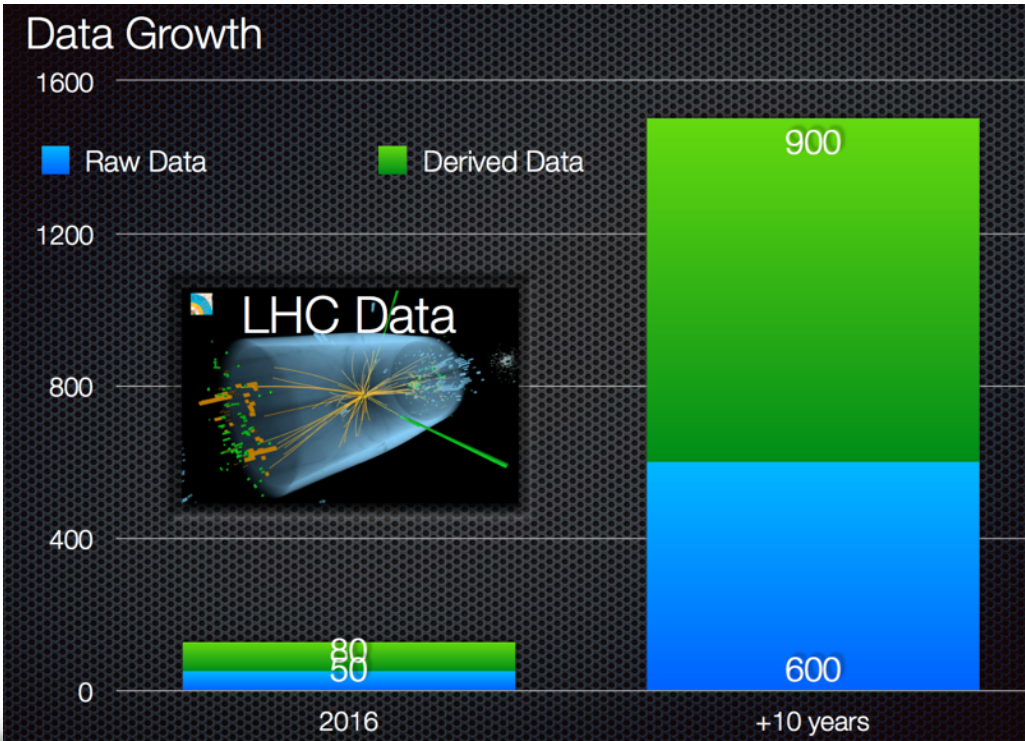
object storage APIs and erasure encoding



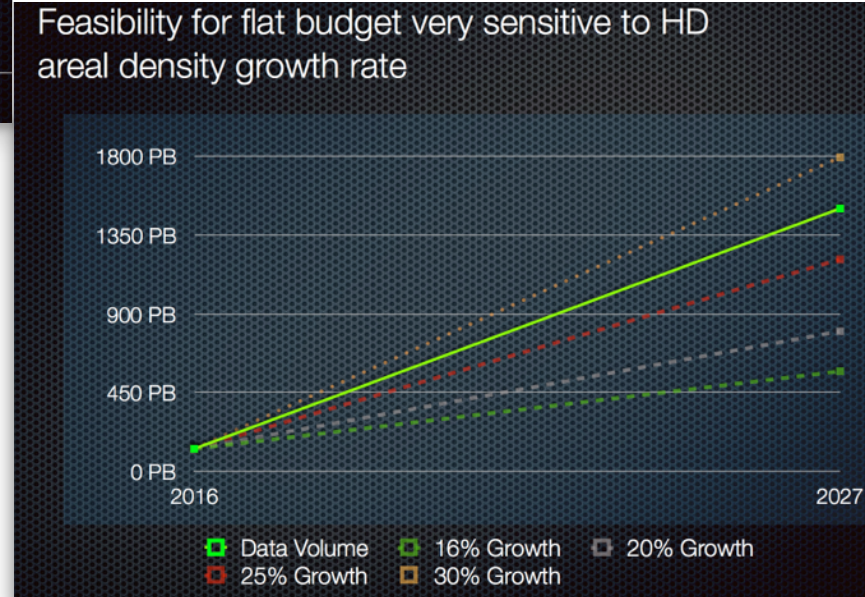
storage tiering



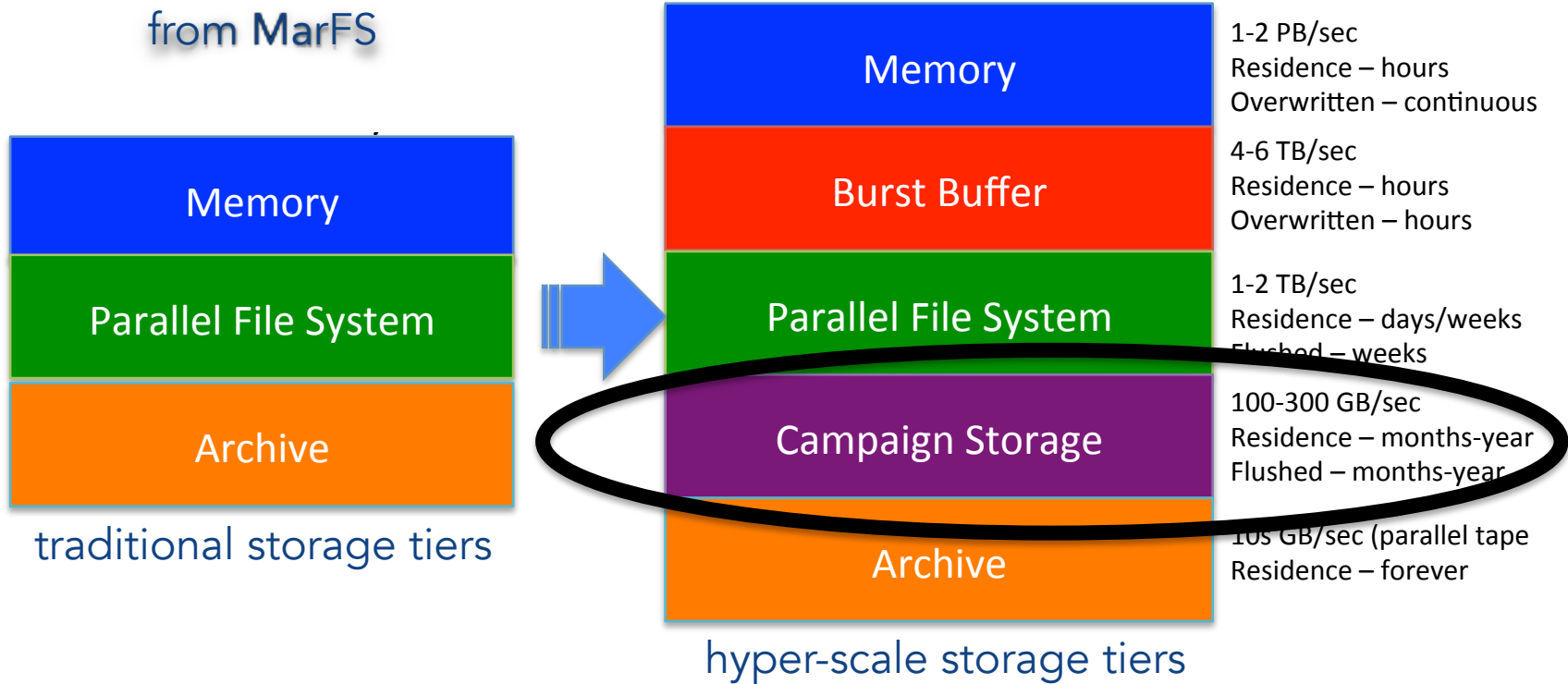
Storage outlook for HL-LHC



we need to optimise storage globally
 need to keep event collection model
 (don't ship or store everything as an event - there are too many)



Vision of Big Data Storage & Federation for HL-LHC



Vision of Big Data Storage & Federation

for HL-LHC

- aim to implement a global federation model for HL-LHC which is **cost-effective** as **campaign storage** system
- **federation of distributed object stores** for data authorised with pre-signed URLs
 - deploy dataloss-free scale-out storage systems with erasure coding ($M, \geq 2$) in big sites
 - data durability should still be guaranteed
 - with global disk replica count = 1
 - with offline replica on cold storage tier
 - compatible with idea of eventually buying public cloud storage
 - model does not exclude to use distributed filesystem as object stores
 - **central scale-out MD & DDM**
 - use few managed storage tiers for global DM, hide small tiers in local federations





Open Source Storage



Open Source Storage



Thank You! Andreas.Joachim.Peters@cern.ch

a reed-solomon code ...



EOS need-to-know info

- **support**
eos-support@cern.ch
- **documentation**
 - <http://eos.readthedocs.io> [en/citrine]
- **repositories**
 - <https://gitlab.cern.ch/dss/eos.git> <https://github.com/cern-eos/eos>
- **mailinglist**
 - eos-community@cern.ch [via egroups.cern.ch]