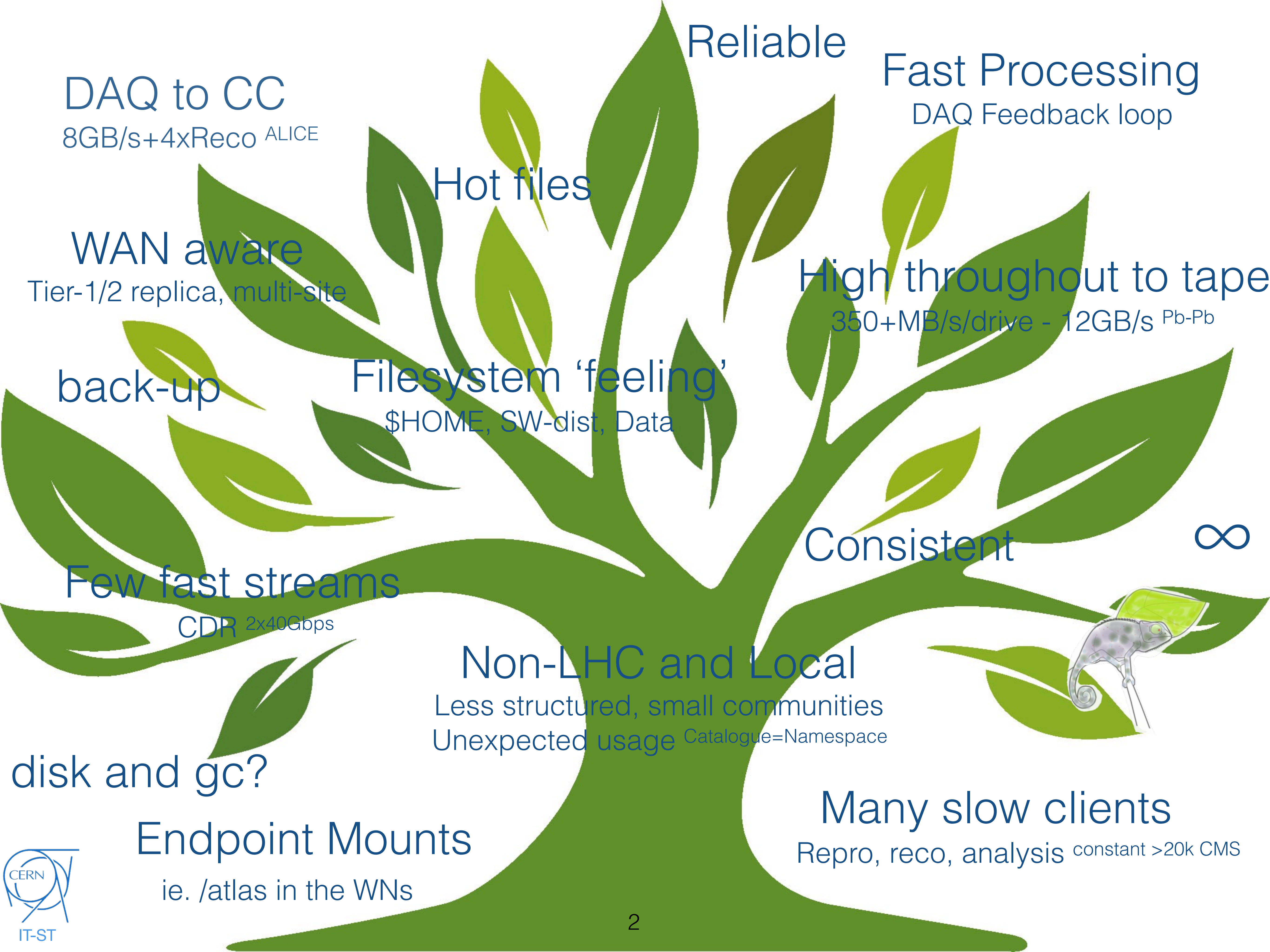
# Storage Chamäleons

2016 Hepix Spring
DESY, Zeuthen

Xavier Espinal
on behalf of IT/ST

25th Anniversary
HEPiX
Zeuthen Spring 2016



DAQ to CC

8GB/s+4xReco ALICE

Reliable

Fast Processing

DAQ Feedback loop

Hot files

WAN aware

Tier-1/2 replica, multi-site

High throughout to tape

350+MB/s/drive - 12GB/s Pb-Pb

back-up

Filesystem ‘feeling’

$HOME, SW-dist, Data

Consistent

∞

Few fast streams

CDR 2x40Gbps

Non-LHC and Local

Less structured, small communities

Unexpected usage Catalogue=Namespace

disk and gc?

Endpoint Mounts

ie. /atlas in the WNs

Many slow clients

Repro, reco, analysis constant >20k CMS

CASTOR
CERN Advanced STORage manager

# Core Systems

Evolved to

## Tape oriented system

Biggest scientific-repo worldwide 138PB and +500M files

High throughput from DAQ, high throughput to tape

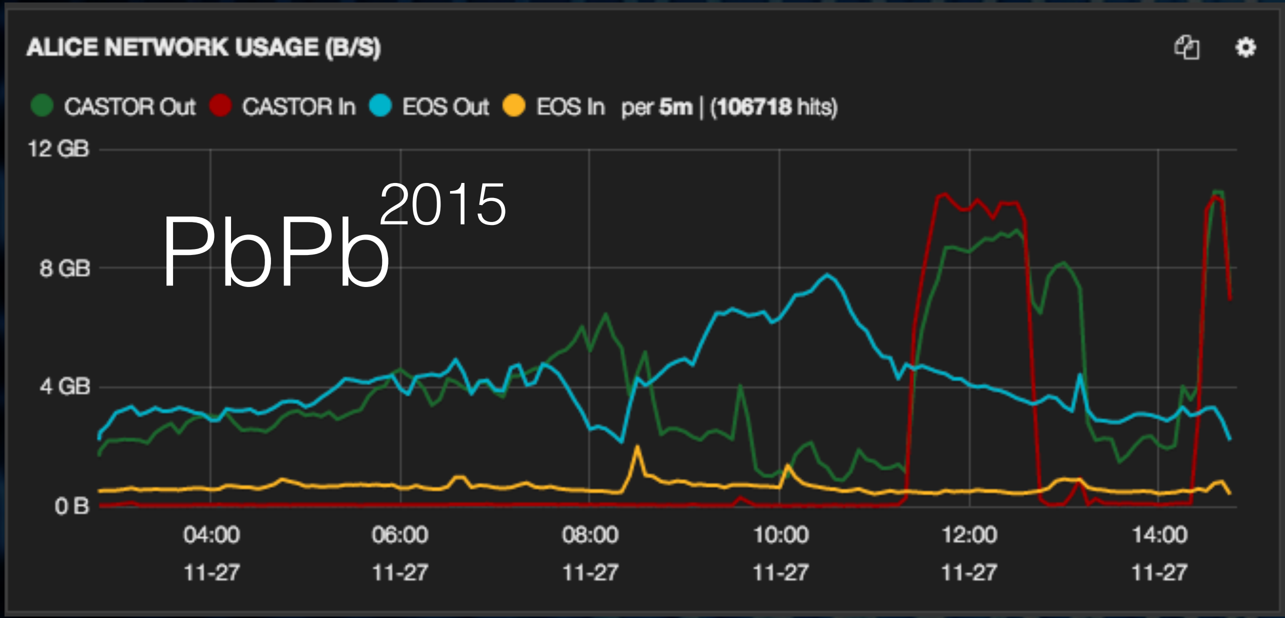
Key feature

## Per stream speed

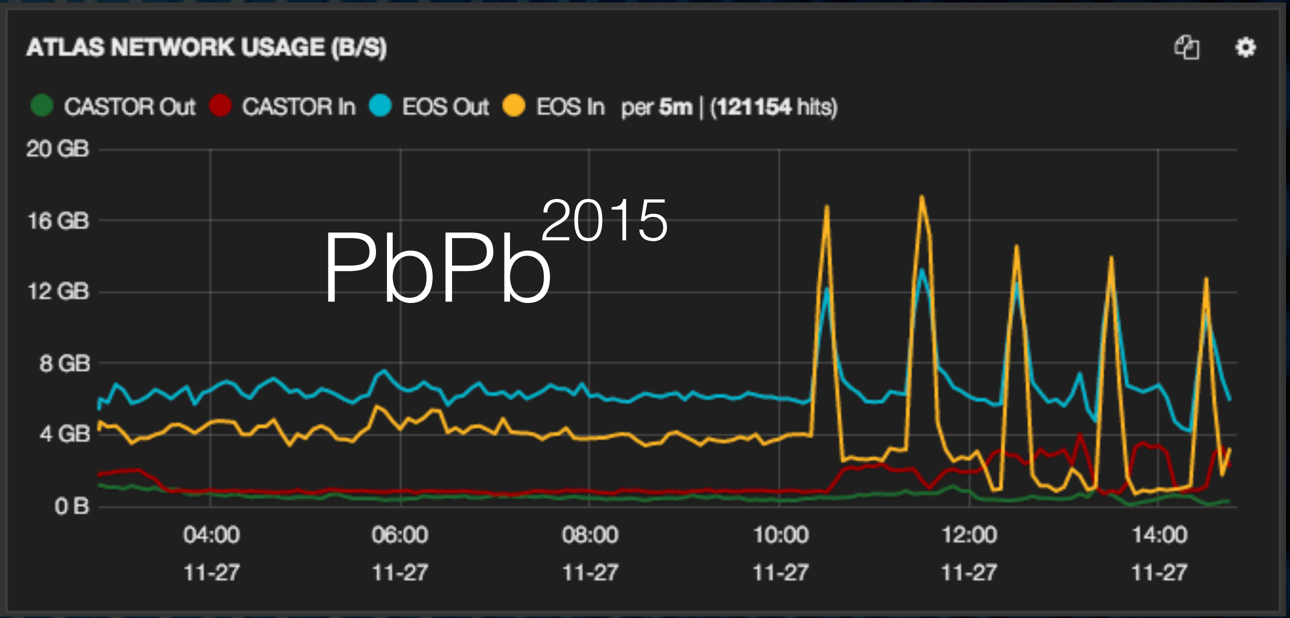
Moved from Raid1 to Raid60 (100MB/s to >350MB/s$^{\text{per stream}}$)

Evaluating common disk layer

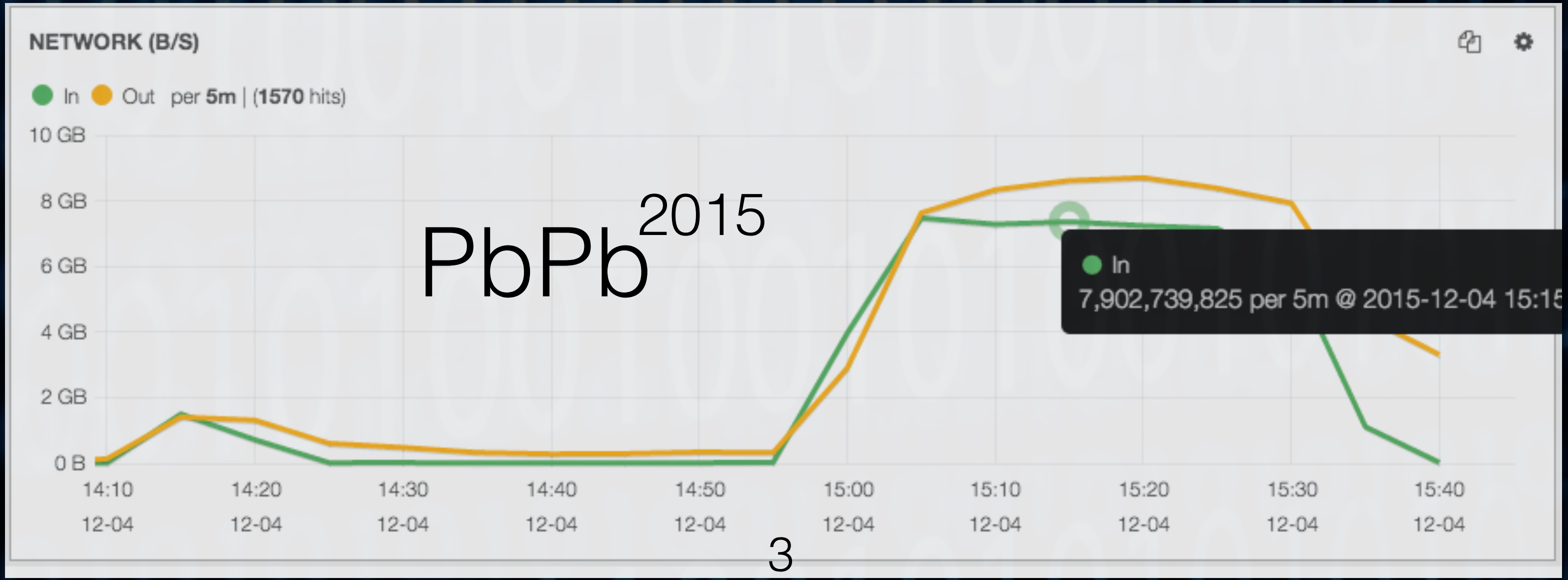
Tape policies, per experiment/user/group resources



PbPb$^{2015}$



PbPb$^{2015}$



PbPb$^{2015}$

CERN
IT-ST

CASTOR
CERN Advanced STORage manager

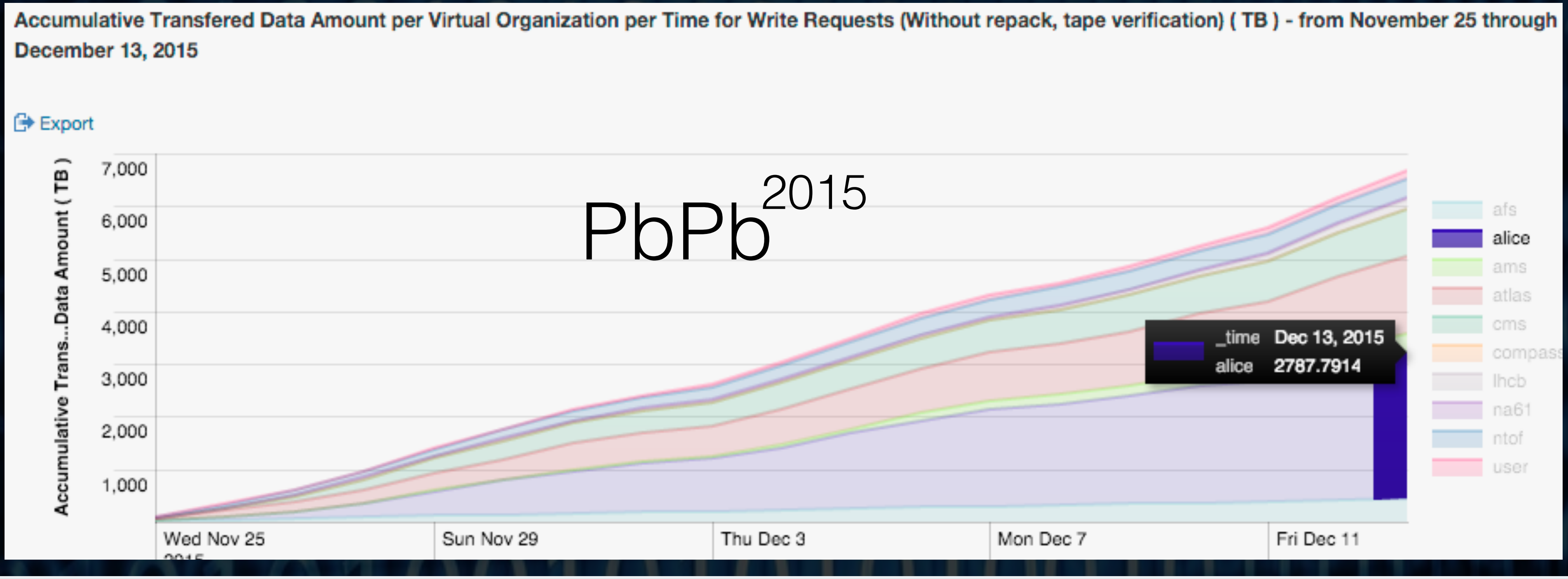
# Core Systems

Evolved to
## Tape oriented system

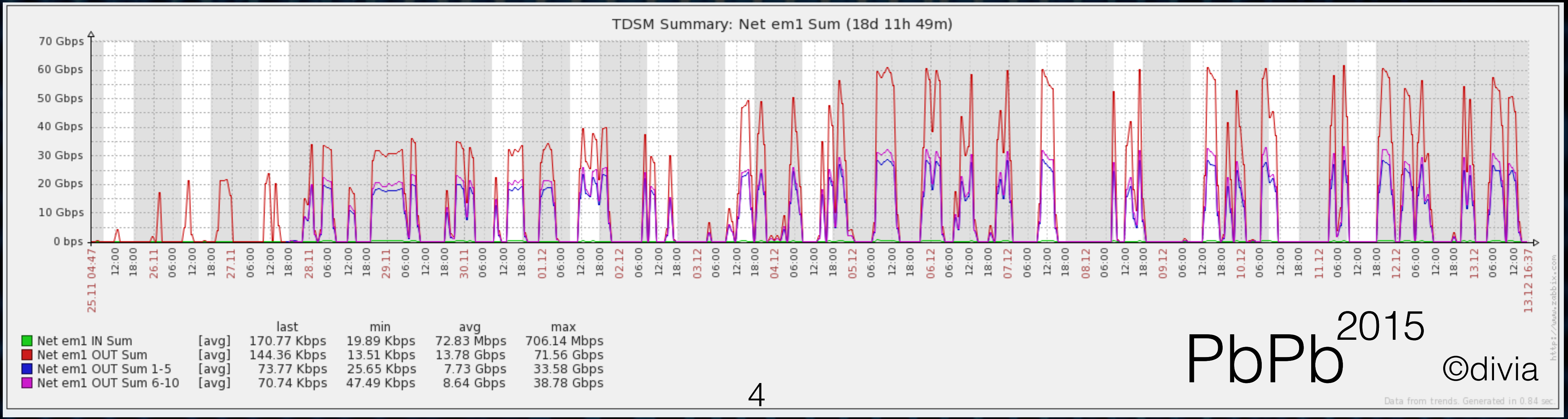
Biggest scientific-repo worldwide 138PB and +500M files
High throughput from DAQ, high throughput to tape

Key feature
## Per stream speed

Moved from Raid1 to Raid60 (100MB/s to >350MB/s$^{\text{per stream}}$)
Evaluating common disk layer
Tape policies, per experiment/user/group resources


**Accumulative Transfered Data Amount per Virtual Organization per Time for Write Requests (Without repack, tape verification) ( TB ) - from November 25 through December 13, 2015**

PbPb$^{2015}$


TDSM Summary: Net em1 Sum (18d 11h 49m)

| | | last | min | avg | max |
|---|---|---|---|---|---|
| 🟩 Net em1 IN Sum | [avg] | 170.77 Kbps | 19.89 Kbps | 72.83 Mbps | 706.14 Mbps |
| 🟥 Net em1 OUT Sum | [avg] | 144.36 Kbps | 13.51 Kbps | 13.78 Gbps | 71.56 Gbps |
| 🟦 Net em1 OUT Sum 1-5 | [avg] | 73.77 Kbps | 25.65 Kbps | 7.73 Gbps | 33.58 Gbps |
| 🟪 Net em1 OUT Sum 6-10 | [avg] | 70.74 Kbps | 47.49 Kbps | 8.64 Gbps | 38.78 Gbps |

Data from trends. Generated in 0.84 sec

PbPb$^{2015}$

©divia

CERN
IT-ST

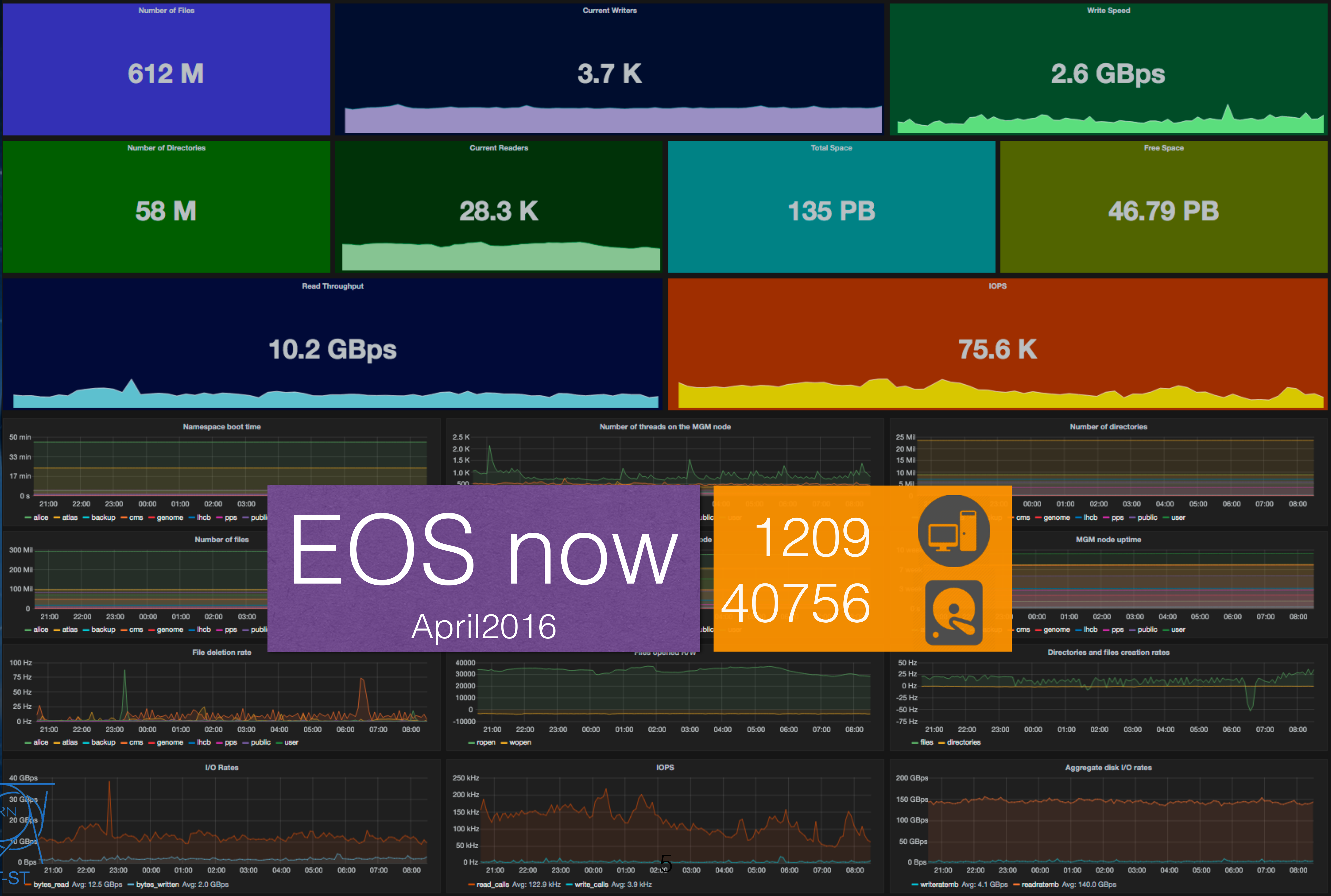
Core Systems

EB era

- Scaling well on #disks
- Performant and manageable
- Main storage platform

NS future

- Fast and consistent
- Horizontally scalable (no single box limitiation)
- zero boot time



Number of Files: 612 M

Current Writers: 3.7 K

Write Speed: 2.6 GBps

Number of Directories: 58 M

Current Readers: 28.3 K

Total Space: 135 PB

Free Space: 46.79 PB

Read Throughput: 10.2 GBps

IOPS: 75.6 K

# EOS now

April2016

1209

40756

CERN IT-ST

# made@CERN

Core Systems

Designed and tailored for experiments needs

Experts in-house: Adapt when required
Re-design if needed

Being used outside: Fermilab, Russia-T1, EsNET, Aarnet
JRC, Univ. Vienna, INFN-Trieste

Openlab/COMTRADE

Adaptable

Catering with different uses

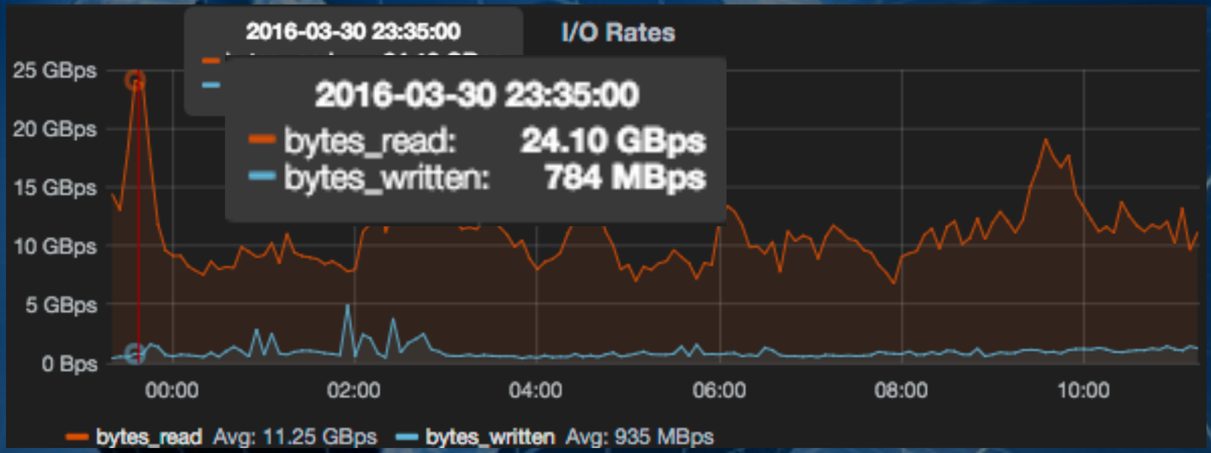
CDR
Data processing
User Analysis
CernBOX

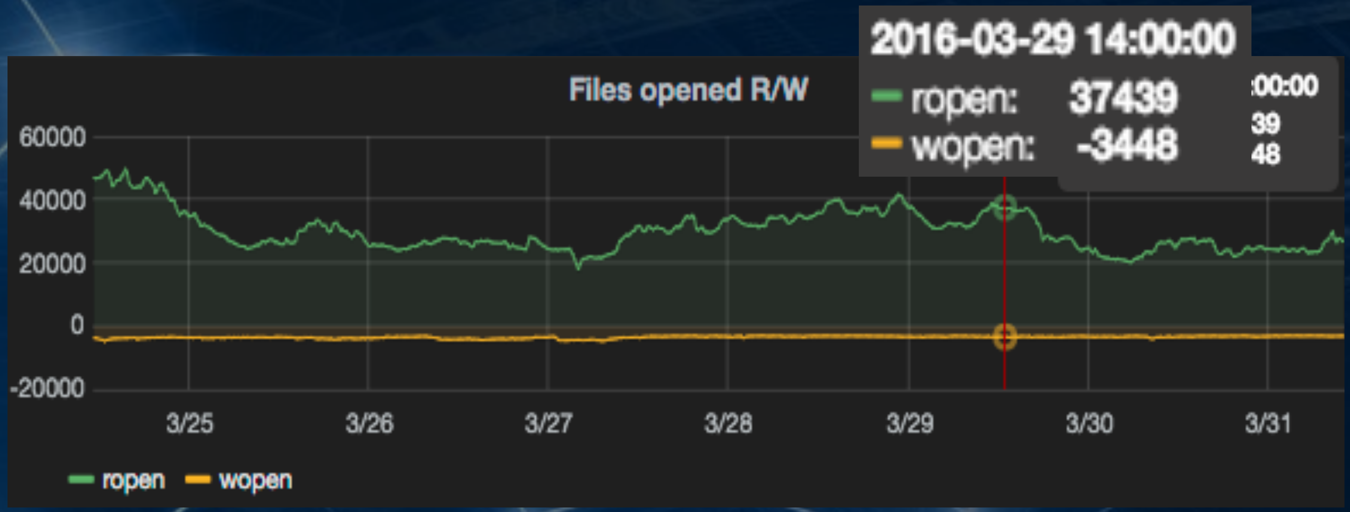
Community data

DmaaS (iJupyter)
Sync

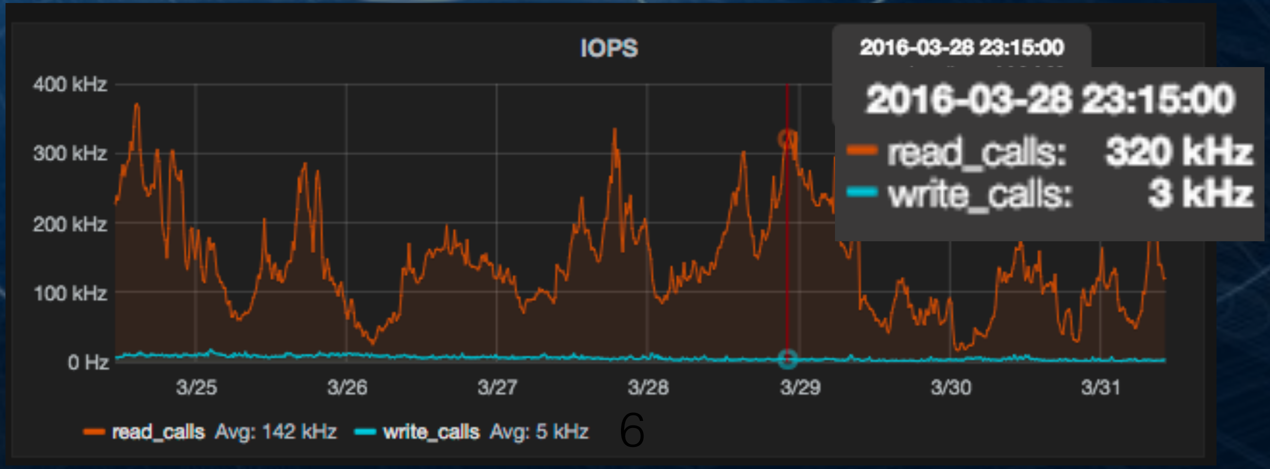
share

CERNBox





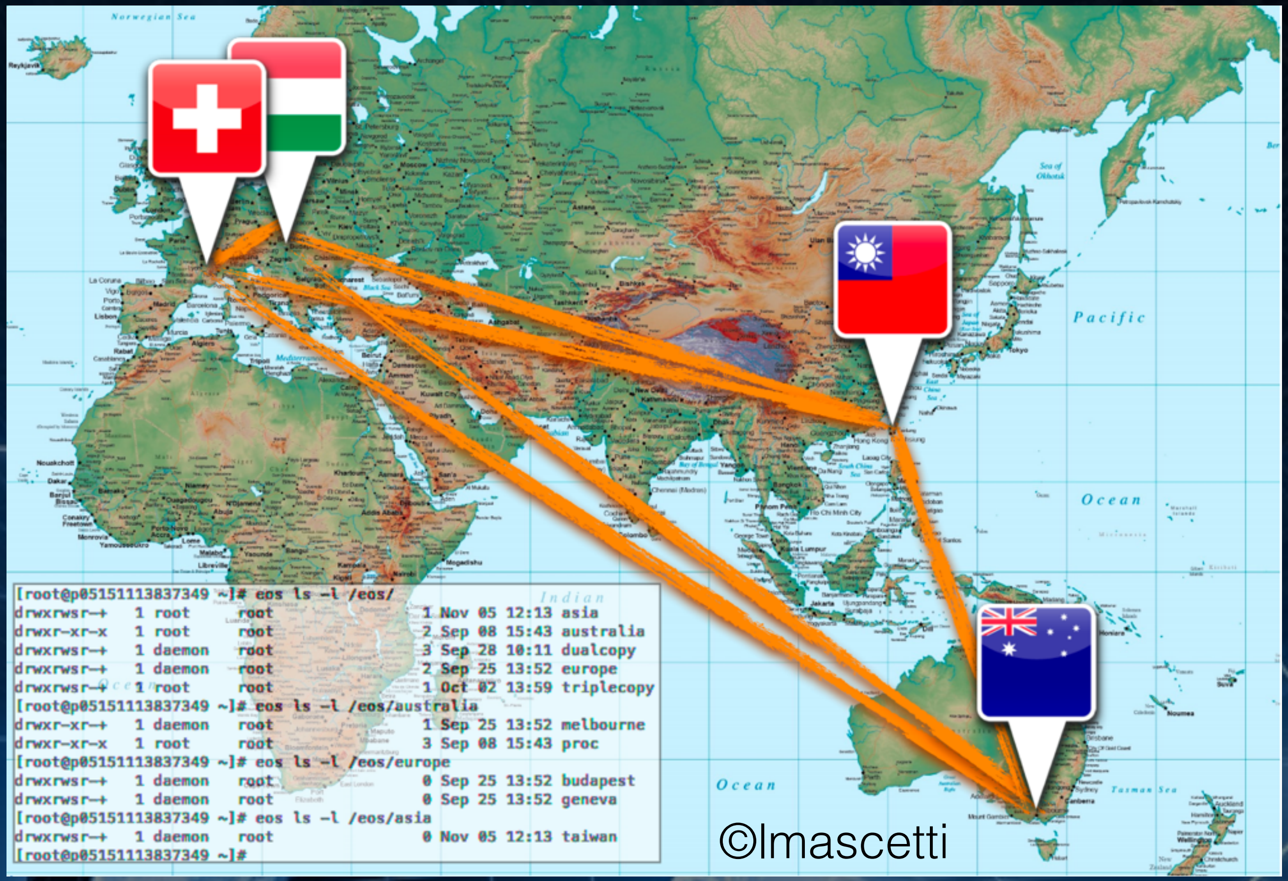
Future Shared FS ? more later…



IT-ST

# Can go distributed, can be shared and synced

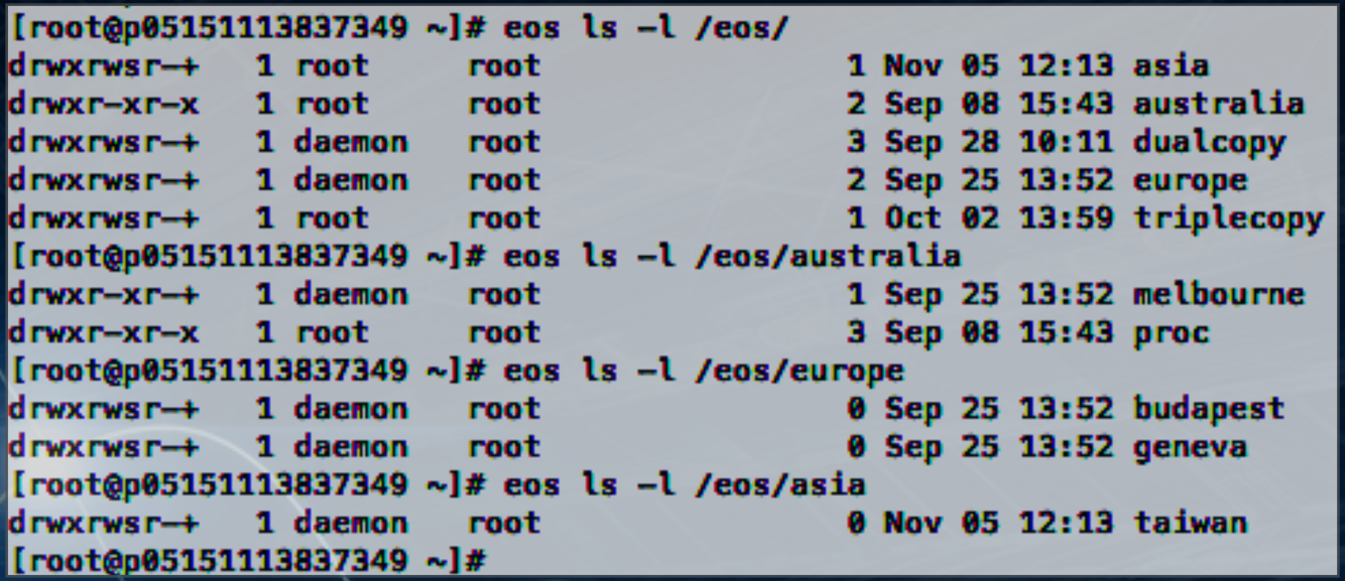
Core Systems


©lmascetti

Clients delocalization

Multi-site deployment

Used from 22 ms to 300 ms



```
[root@p05151113837349 ~]# eos ls -l /eos/
drwxrwsr-+   1 root     root              1 Nov 05 12:13 asia
drwxr-xr-x   1 root     root              2 Sep 08 15:43 australia
drwxrwsr-+   1 daemon   root              3 Sep 28 10:11 dualcopy
drwxrwsr-+   1 daemon   root              2 Sep 25 13:52 europe
drwxrwsr-+   1 root     root              1 Oct 02 13:59 triplecopy
[root@p05151113837349 ~]# eos ls -l /eos/australia
drwxr-xr-+   1 daemon   root              1 Sep 25 13:52 melbourne
drwxr-xr-x   1 root     root              3 Sep 08 15:43 proc
[root@p05151113837349 ~]# eos ls -l /eos/europe
drwxrwsr-+   1 daemon   root              0 Sep 25 13:52 budapest
drwxrwsr-+   1 daemon   root              0 Sep 25 13:52 geneva
[root@p05151113837349 ~]# eos ls -l /eos/asia
drwxrwsr-+   1 daemon   root              0 Nov 05 12:13 taiwan
[root@p05151113837349 ~]#
```

Community data share

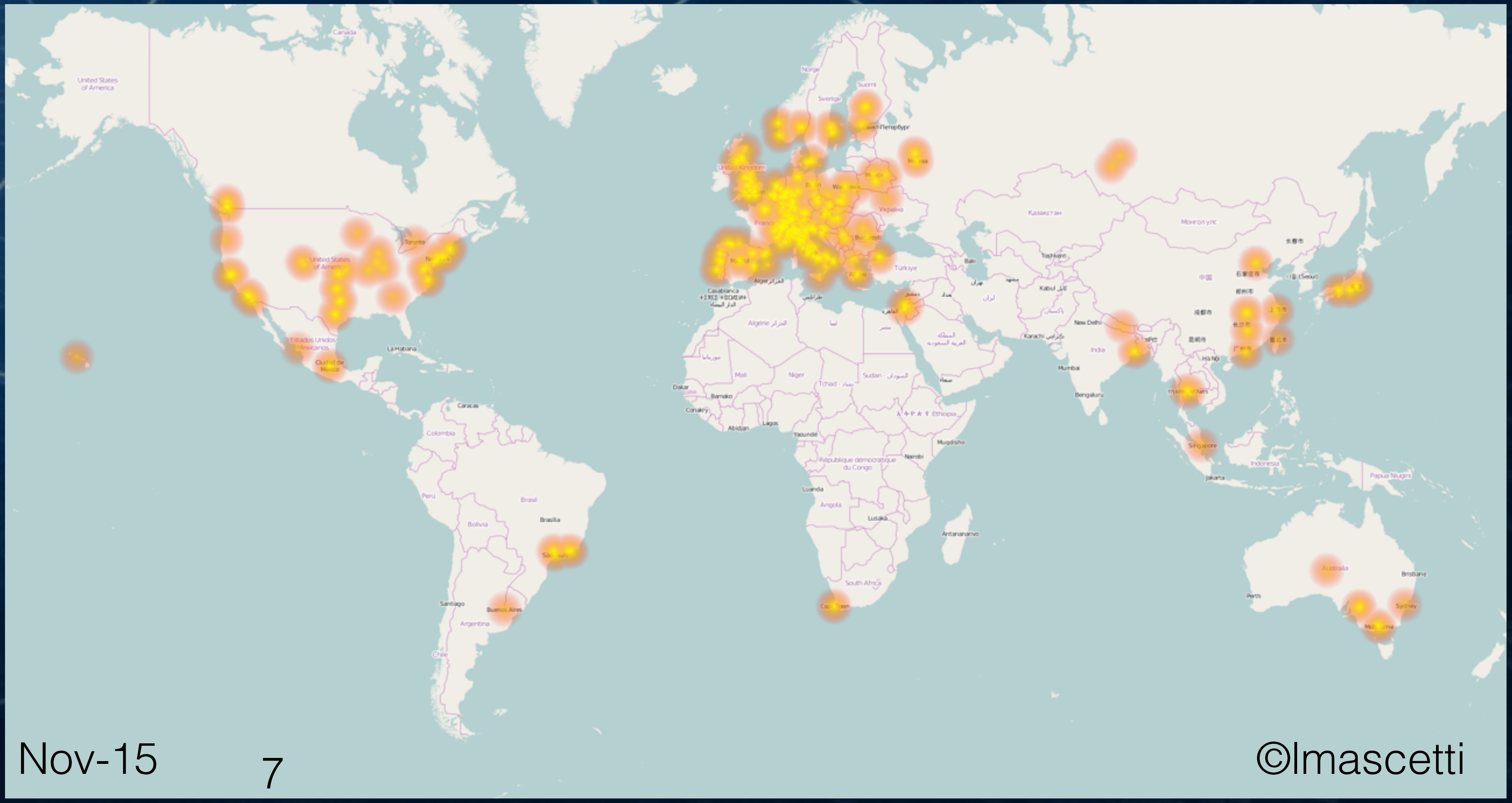
DmaaS (iJupyter)

Sync

CERNBox

| | |
|---|---|
| Users | 4719 |
| # files | 70 Million |
| # dirs | 9 Million |
| Quota | 1TB/user |
| Used Space | 125 TB |
| Deployed Space | 1.5 PB |

20% 60% 20%

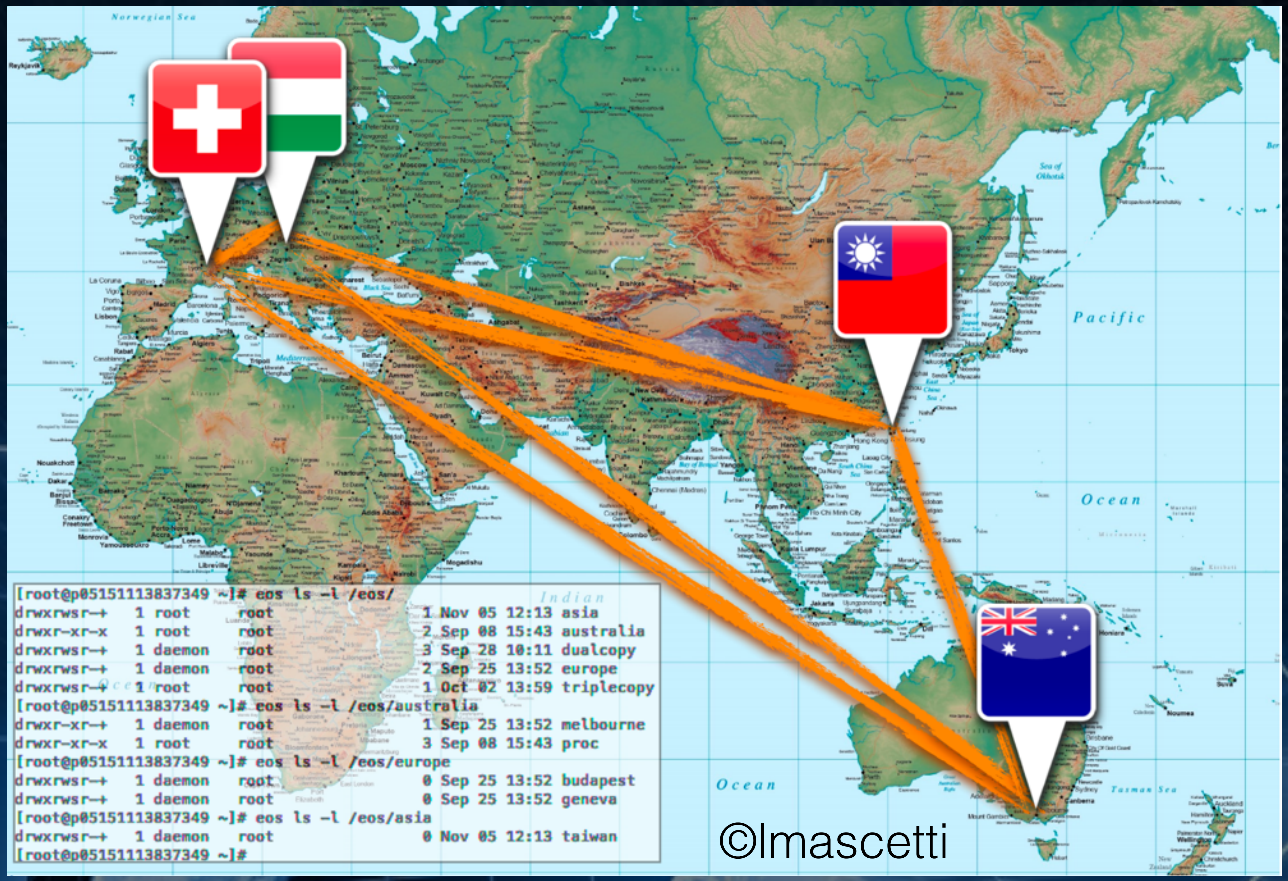

©lmascetti

Nov-15

CERN IT-ST

# Can go distributed, can be shared and synced

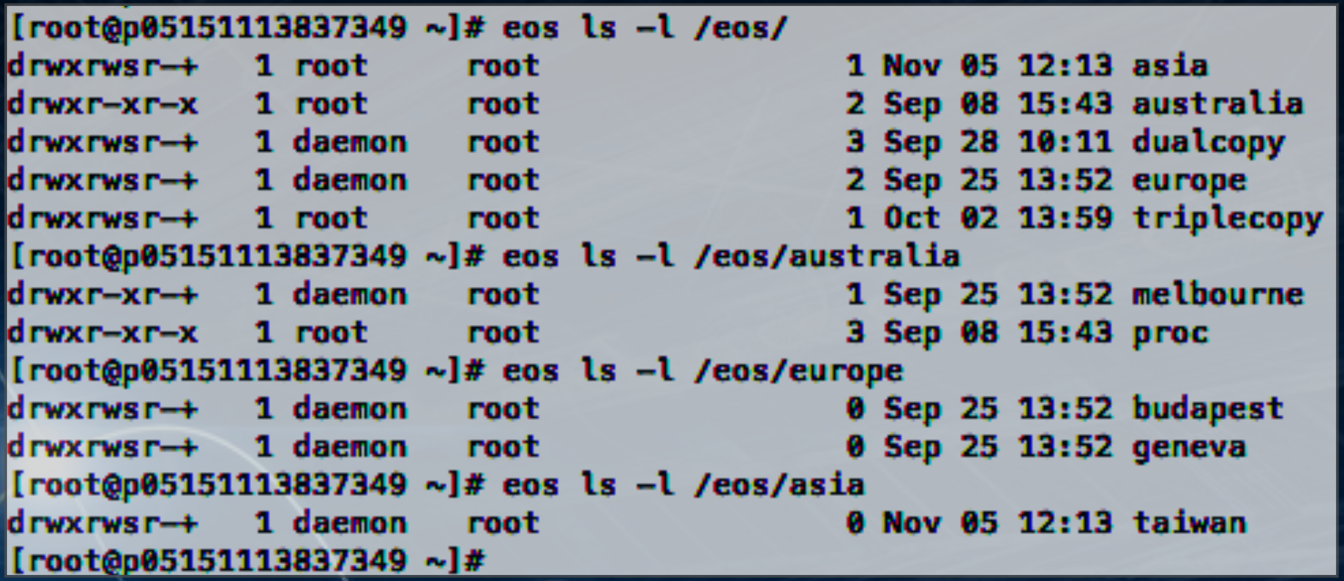
Core Systems


©lmascetti

Clients delocalization

Multi-site deployment

Used from 22 ms to 300 ms



```
[root@p05151113837349 ~]# eos ls -l /eos/
drwxrwsr-+   1 root     root               1 Nov 05 12:13 asia
drwxr-xr-x   1 root     root               2 Sep 08 15:43 australia
drwxrwsr-+   1 daemon   root               3 Sep 28 10:11 dualcopy
drwxrwsr-+   1 daemon   root               2 Sep 25 13:52 europe
drwxrwsr-+   1 root     root               1 Oct 02 13:59 triplecopy
[root@p05151113837349 ~]# eos ls -l /eos/australia
drwxr-xr-+   1 daemon   root               1 Sep 25 13:52 melbourne
drwxr-xr-x   1 root     root               3 Sep 08 15:43 proc
[root@p05151113837349 ~]# eos ls -l /eos/europe
drwxrwsr-+   1 daemon   root               0 Sep 25 13:52 budapest
drwxrwsr-+   1 daemon   root               0 Sep 25 13:52 geneva
[root@p05151113837349 ~]# eos ls -l /eos/asia
drwxrwsr-+   1 daemon   root               0 Nov 05 12:13 taiwan
[root@p05151113837349 ~]#
```

Community data share

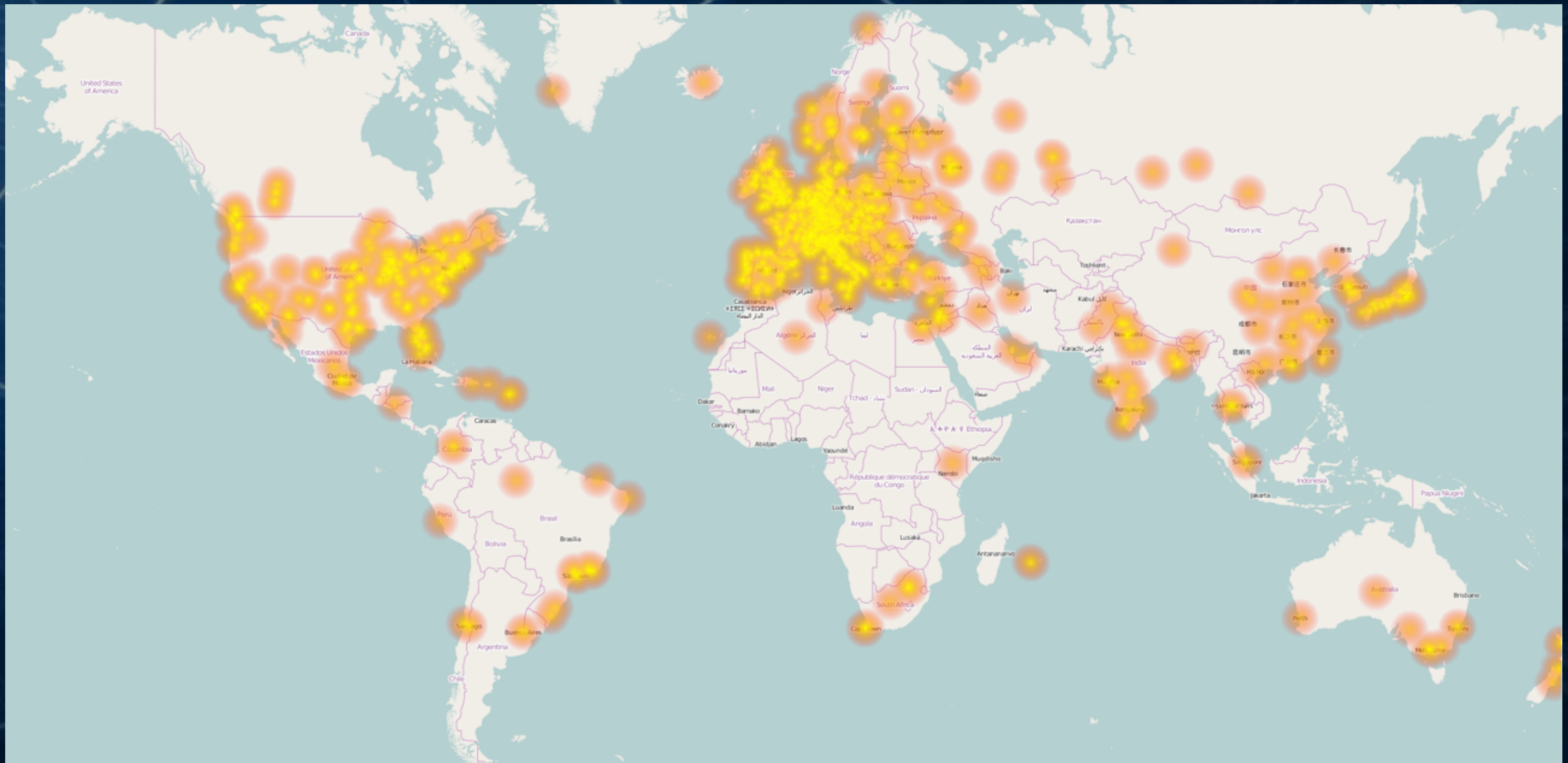
DmaaS (iJupyter)

Sync

CERNBox

| | |
|---|---|
| Users | 4719 |
| # files | 70 Million |
| # dirs | 9 Million |
| Quota | 1TB/user |
| Used Space | 125 TB |
| Deployed Space | 1.5 PB |

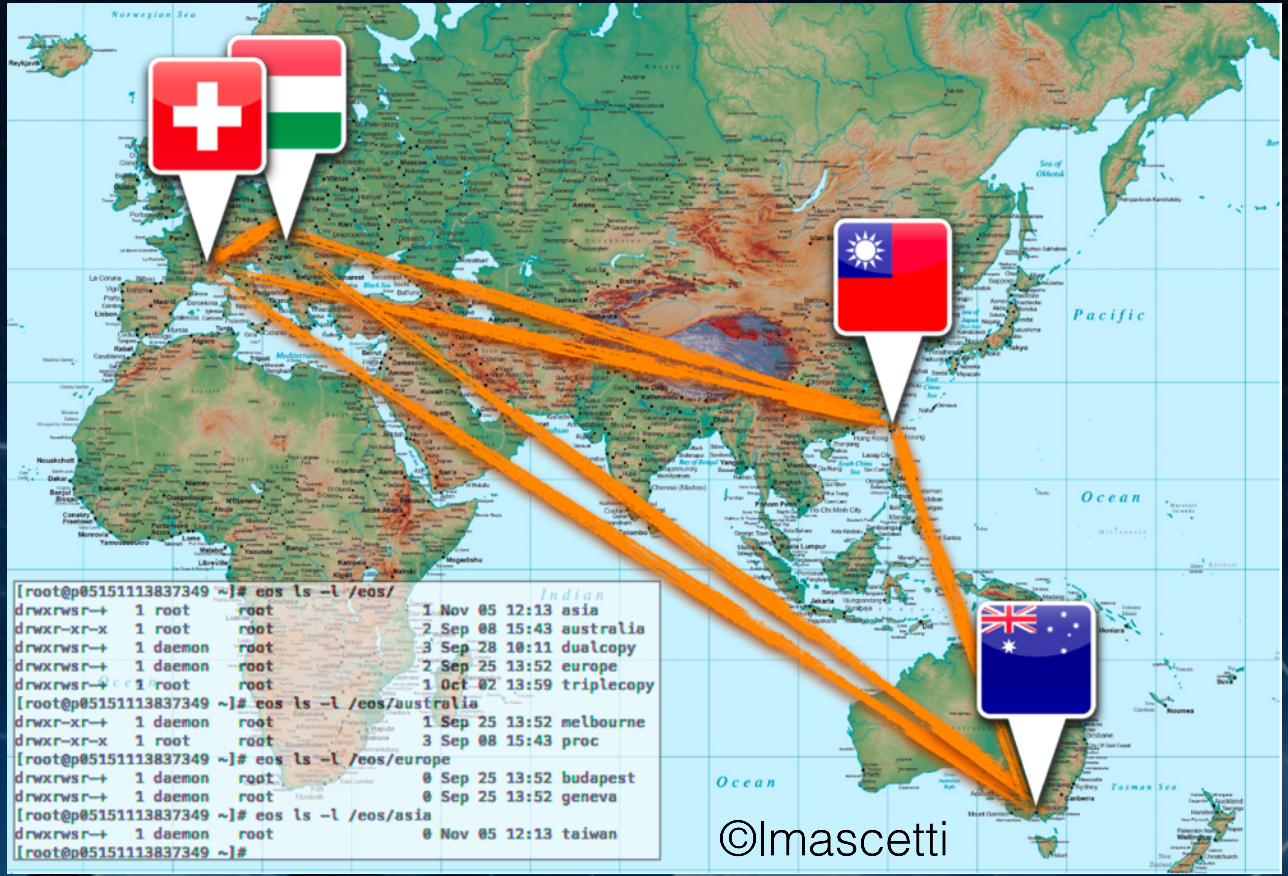
20% 60% 20%


©lmascetti

EOS

# Can go distributed, can be shared and synced

Core Systems



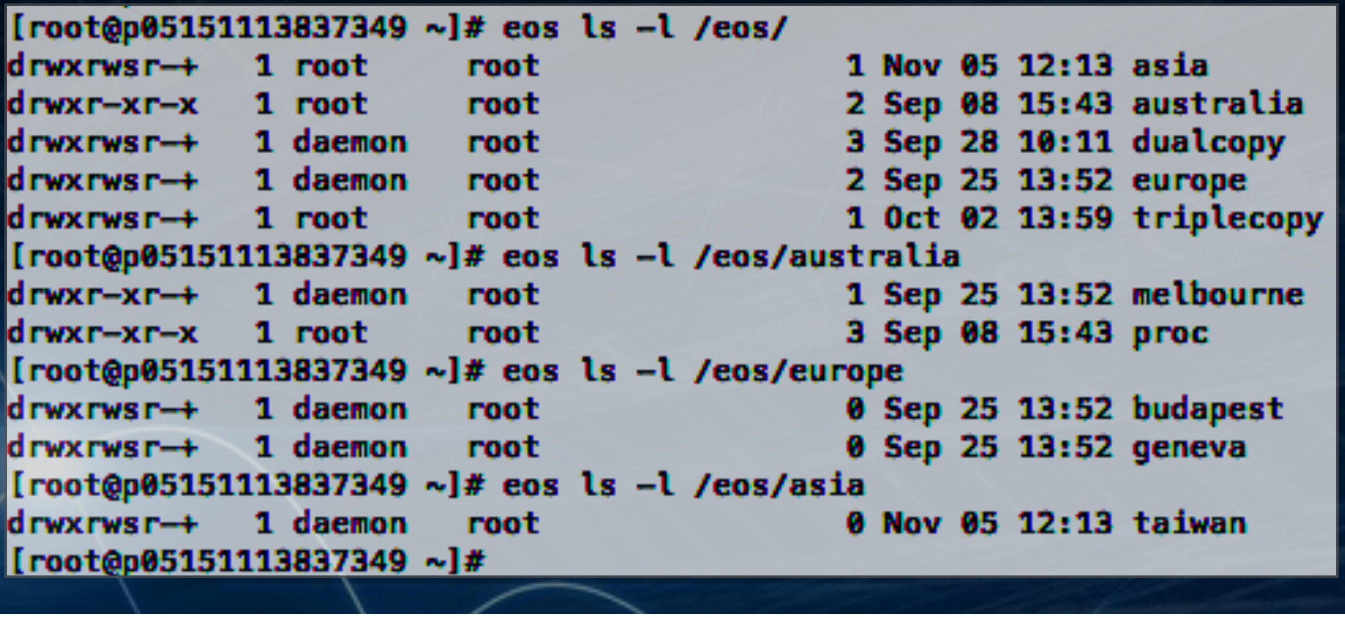
```
[root@p05151113837349 ~]# eos ls -l /eos/
drwxrwsr-+   1 root     root              1 Nov 05 12:13 asia
drwxr-xr-x   1 root     root              2 Sep 08 15:43 australia
drwxrwsr-+   1 daemon   root              3 Sep 28 10:11 dualcopy
drwxrwsr-+   1 daemon   root              2 Sep 25 13:52 europe
drwxrwsr-+   1 root     root              1 Oct 02 13:59 triplecopy
[root@p05151113837349 ~]# eos ls -l /eos/australia
drwxr-xr-+   1 daemon   root              1 Sep 25 13:52 melbourne
drwxr-xr-x   1 root     root              3 Sep 08 15:43 proc
[root@p05151113837349 ~]# eos ls -l /eos/europe
drwxrwsr-+   1 daemon   root              0 Sep 25 13:52 budapest
drwxrwsr-+   1 daemon   root              0 Sep 25 13:52 geneva
[root@p05151113837349 ~]# eos ls -l /eos/asia
drwxrwsr-+   1 daemon   root              0 Nov 05 12:13 taiwan
[root@p05151113837349 ~]#
```

©lmascetti

Clients delocalization

Multi-site deployment

Used from 22 ms to 300 ms



```
[root@p05151113837349 ~]# eos ls -l /eos/
drwxrwsr-+   1 root     root              1 Nov 05 12:13 asia
drwxr-xr-x   1 root     root              2 Sep 08 15:43 australia
drwxrwsr-+   1 daemon   root              3 Sep 28 10:11 dualcopy
drwxrwsr-+   1 daemon   root              2 Sep 25 13:52 europe
drwxrwsr-+   1 root     root              1 Oct 02 13:59 triplecopy
[root@p05151113837349 ~]# eos ls -l /eos/australia
drwxr-xr-+   1 daemon   root              1 Sep 25 13:52 melbourne
drwxr-xr-x   1 root     root              3 Sep 08 15:43 proc
[root@p05151113837349 ~]# eos ls -l /eos/europe
drwxrwsr-+   1 daemon   root              0 Sep 25 13:52 budapest
drwxrwsr-+   1 daemon   root              0 Sep 25 13:52 geneva
[root@p05151113837349 ~]# eos ls -l /eos/asia
drwxrwsr-+   1 daemon   root              0 Nov 05 12:13 taiwan
[root@p05151113837349 ~]#
```

Community data share

DmaaS (iJupyter) Sync

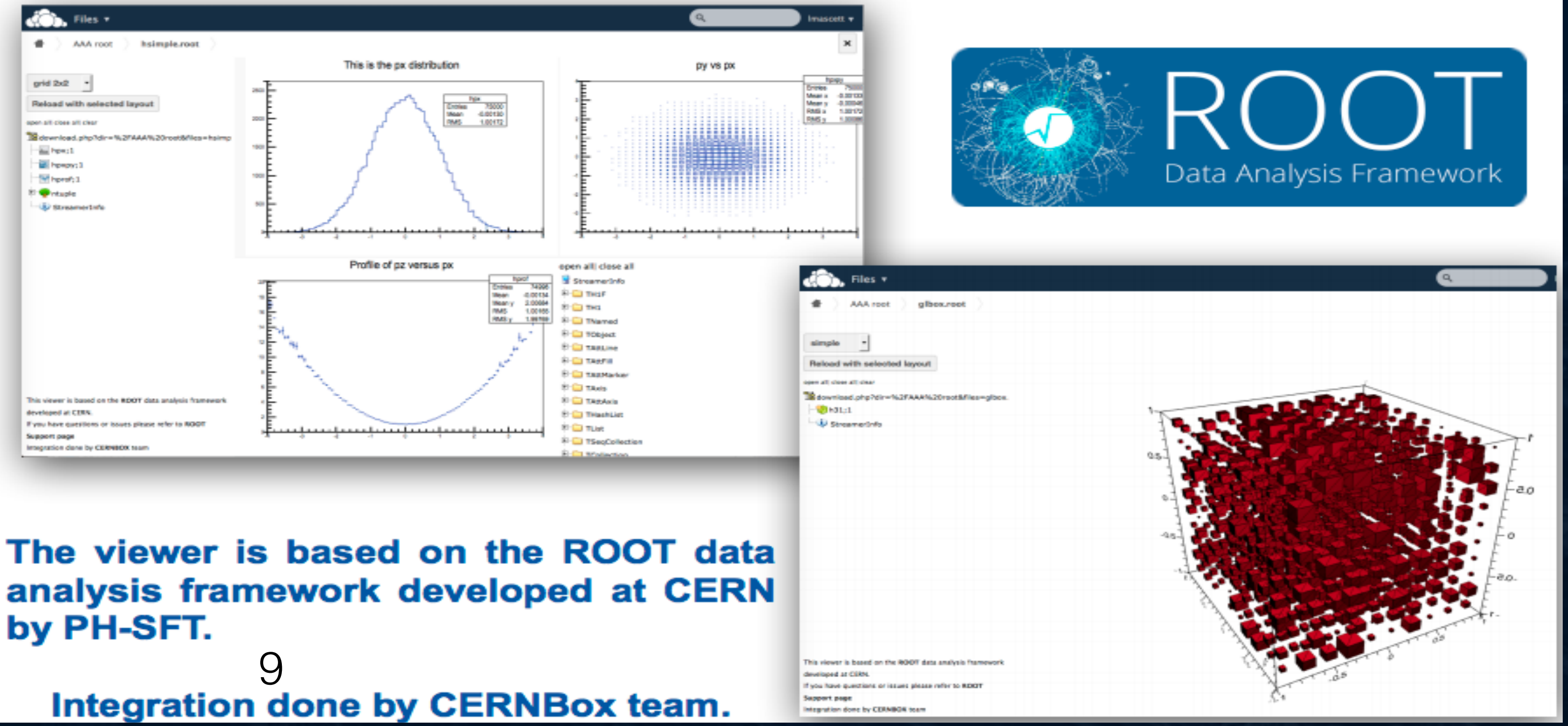
CERNBox

| | |
|---|---|
| Users | 4719 |
| # files | 70 Million |
| # dirs | 9 Million |
| Quota | 1TB/user |
| Used Space | 125 TB |
| Deployed Space | 1.5 PB |

20% 60% 20%

## Embedded ROOT Viewer

©dpiparo



This viewer is based on the ROOT data analysis framework developed at CERN.
If you have questions or issues please refer to ROOT
Support page
Integration done by CERNBOX team

**The viewer is based on the ROOT data analysis framework developed at CERN by PH-SFT.**

**Integration done by CERNBox team.**

# BLOCK STORAGE

## Core Systems

Openstack VM

Cinder Volumes

S3

RADOS FS

File stripper

CASTOR backend

Under evaluation

Large contribution

Community

Code development

CERN-IT/ST

Largest Cluster 30PB

Deployed to date

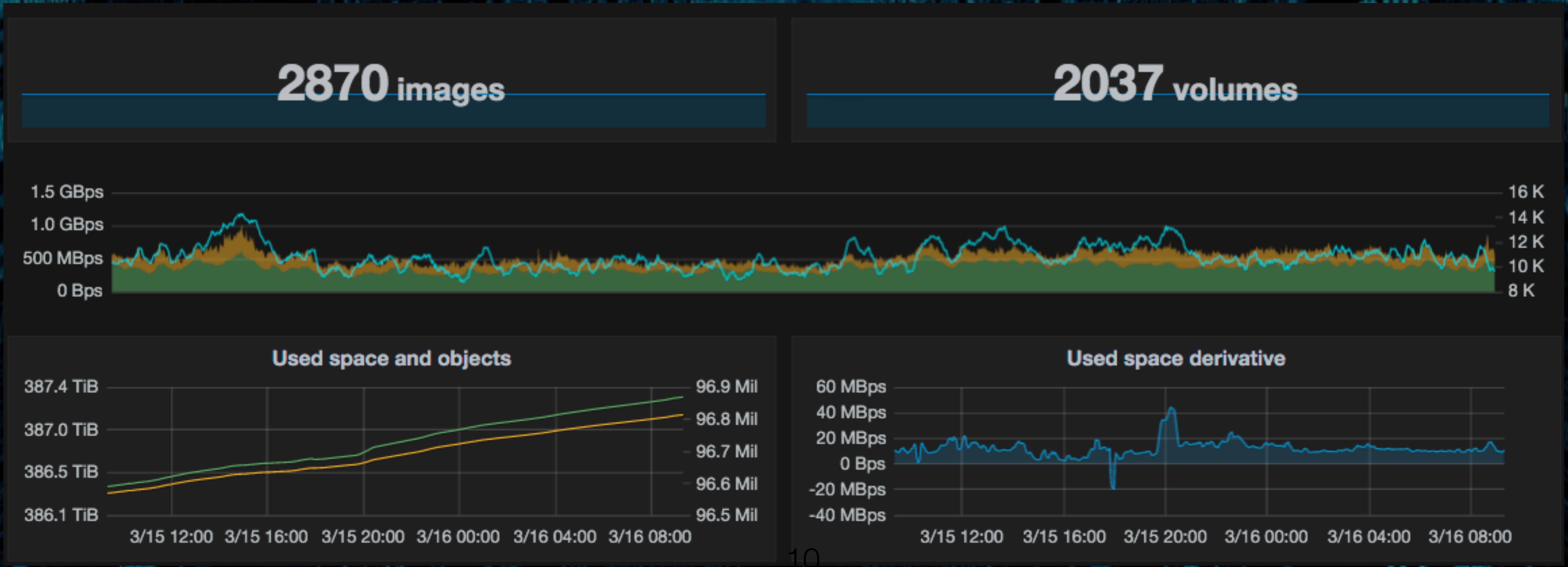
40k OSDs

Multi-site

In production

3PB@wigner

1PB@meyrin



2870 images

2037 volumes

Used space and objects

Used space derivative

IT-ST

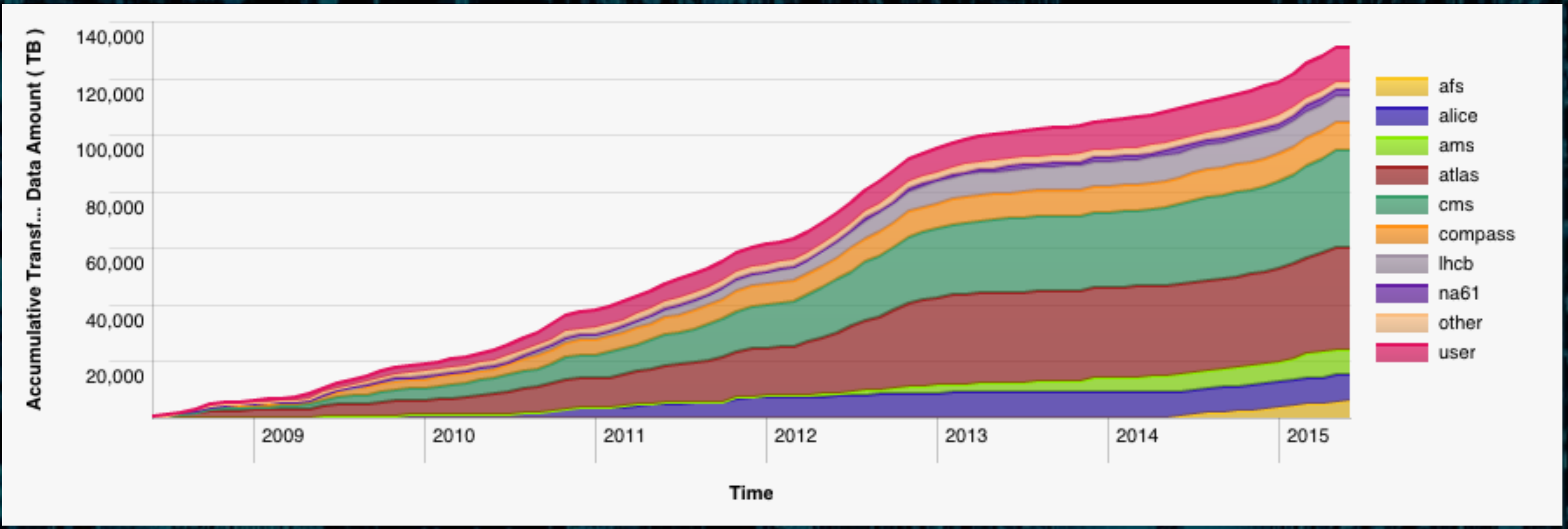
# CERN Tape Archive

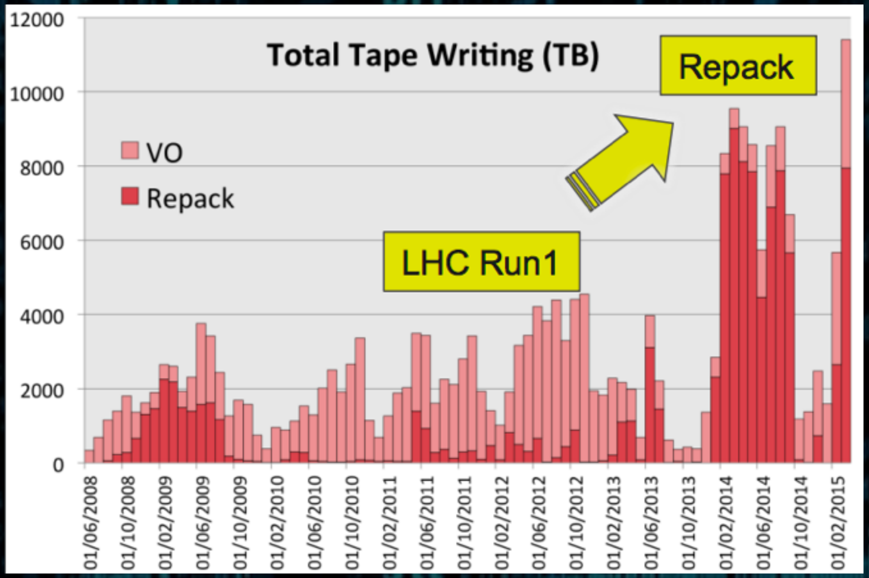
Technology driven: new medias brings ↑density ↑speed

Towards a pluggable tape backend (EOS)

Cold by definition: hight throughput, high latency

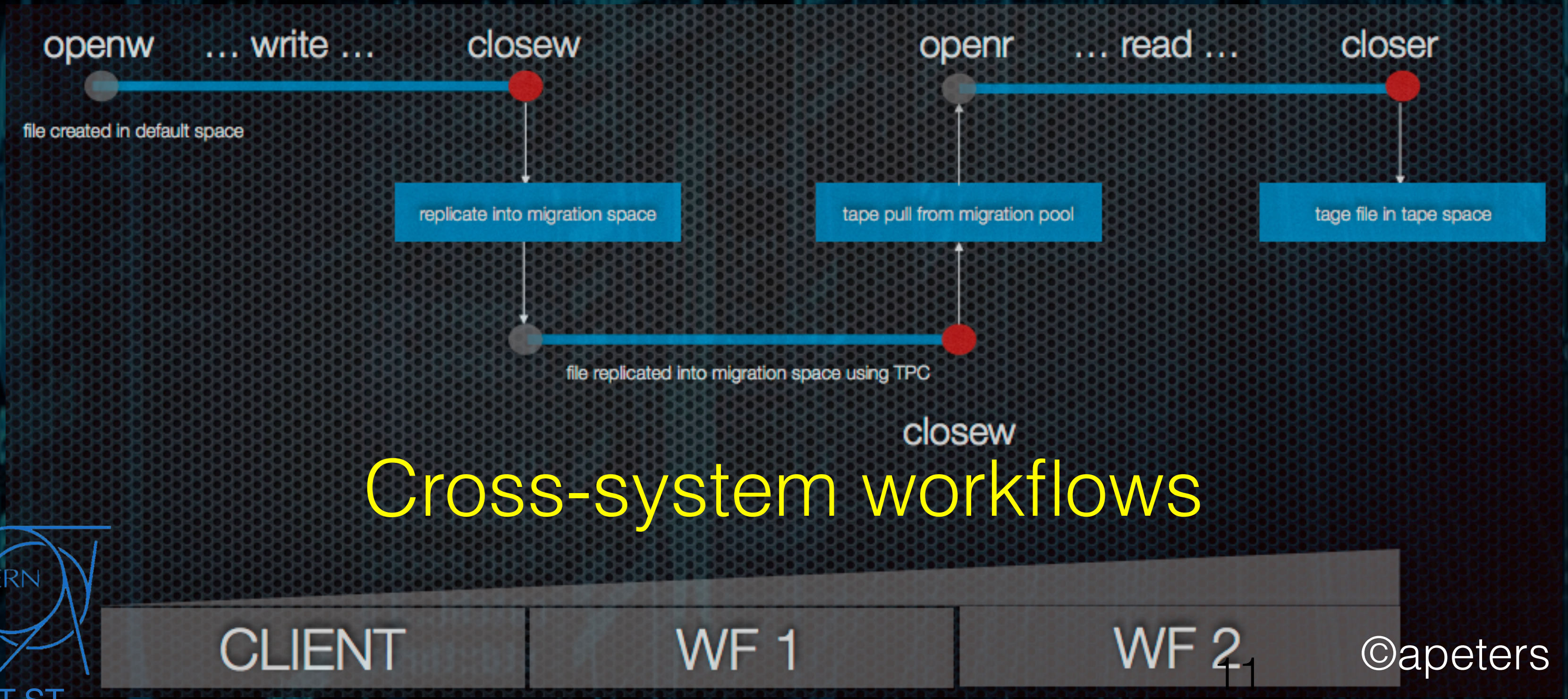
Core Systems





Tape best technology for data repositories: TCO media power density and resilient/reliable

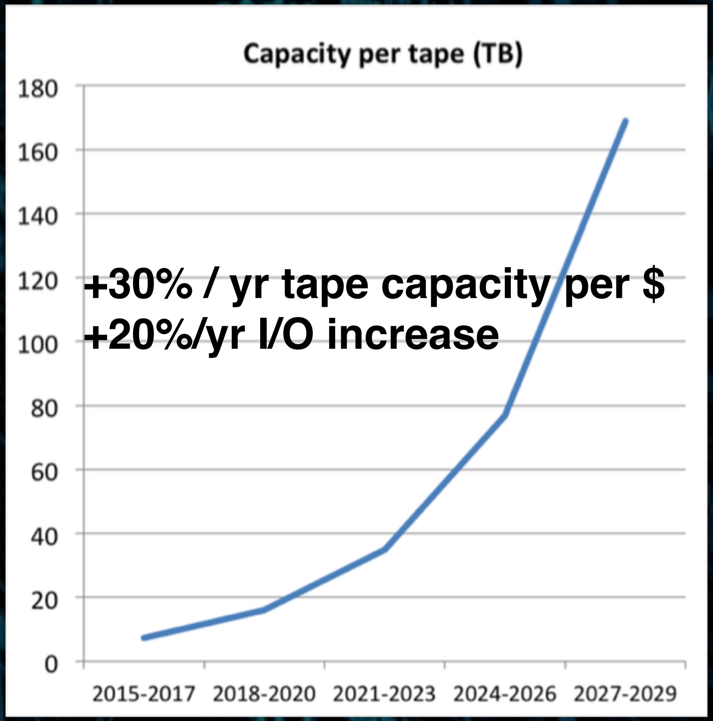
very large disk caches nowadays



openw
... write ...
closew
file created in default space
replicate into migration space
file replicated into migration space using TPC
openr
... read ...
closer
tape pull from migration pool
tage file in tape space
closew
CLIENT
WF 1
WF 2

Cross-system workflows

©apeters



**+30% / yr tape capacity per $**

**+20%/yr I/O increase**

CERN

IT-ST

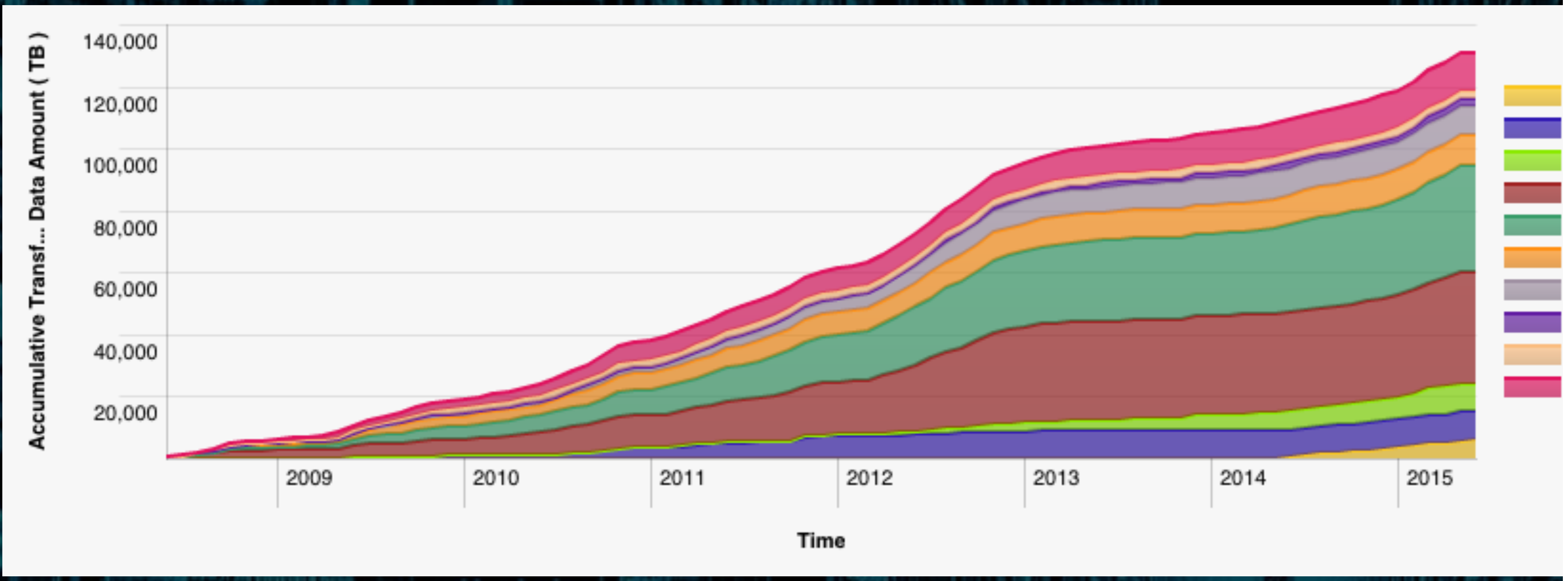
# CERN Tape Archive

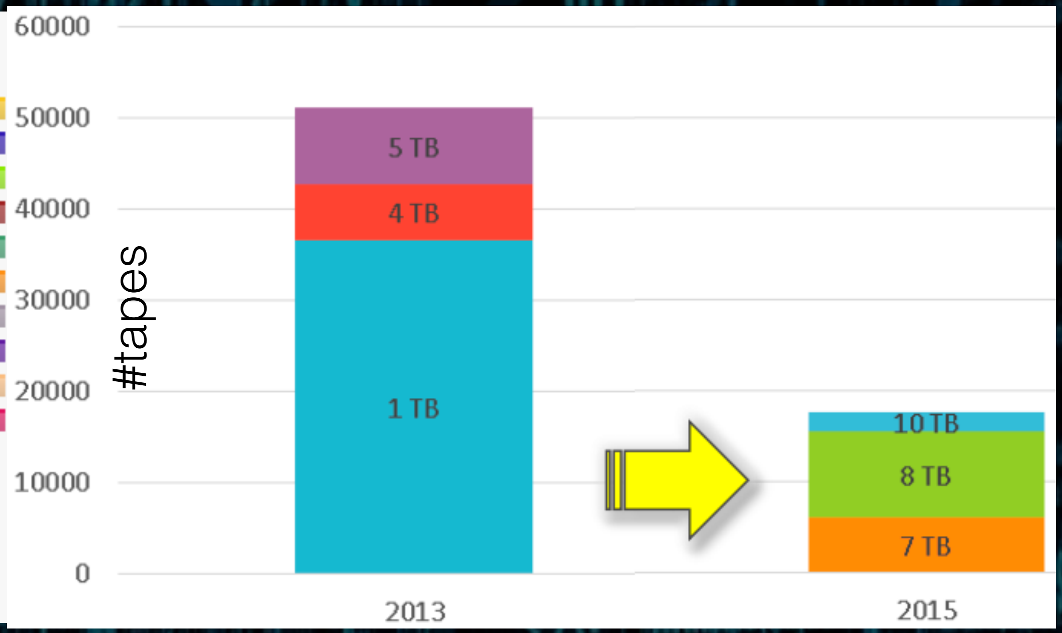
Technology driven: new medias brings ↑ density ↑ speed

Towards a pluggable tape backend (EOS)

Cold by definition: hight throughput, high latency

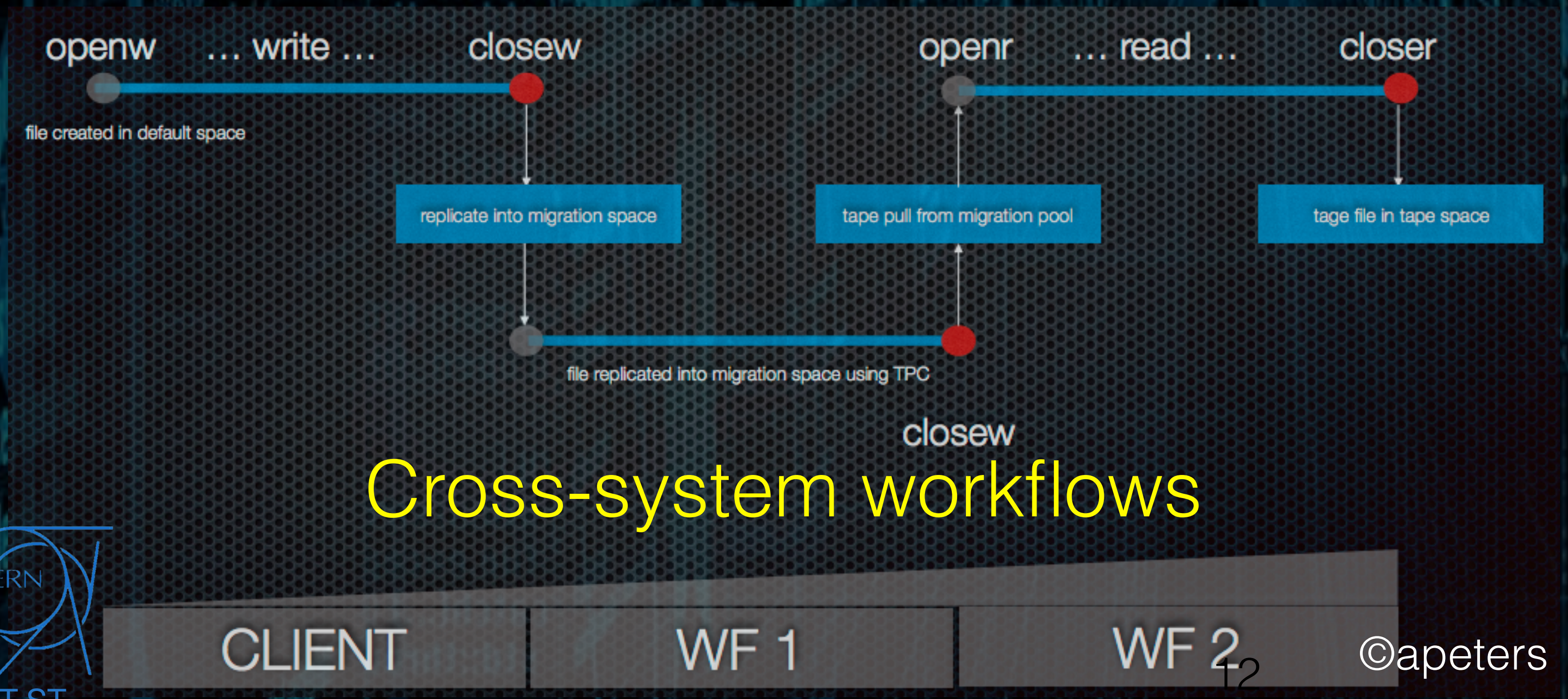
Core Systems





Tape best technology for data repositories: TCO media power density and resilient/reliable

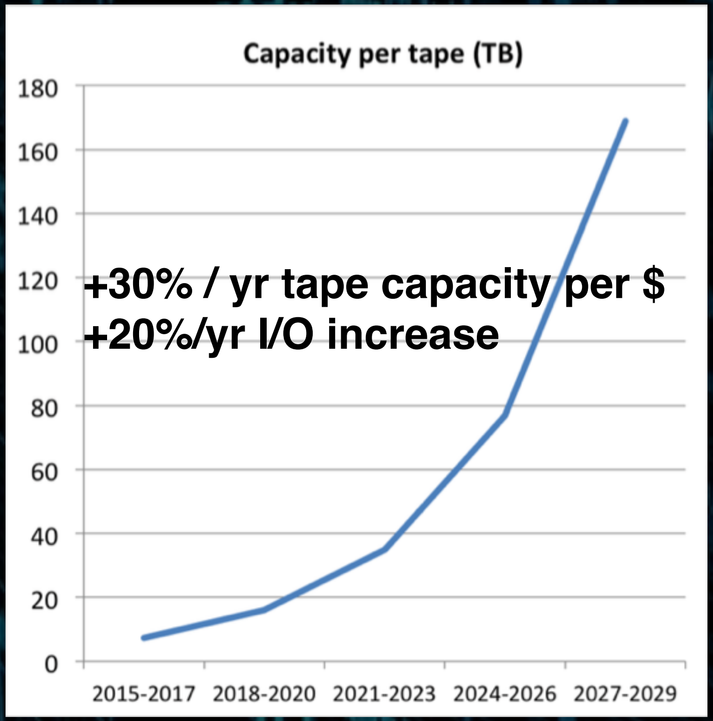
very large disk caches nowadays



openw
... write ...
closew
openr
... read ...
closer
file created in default space
replicate into migration space
tape pull from migration pool
tage file in tape space
file replicated into migration space using TPC
closew
CLIENT
WF 1
WF 2

Cross-system workflows

©apeters



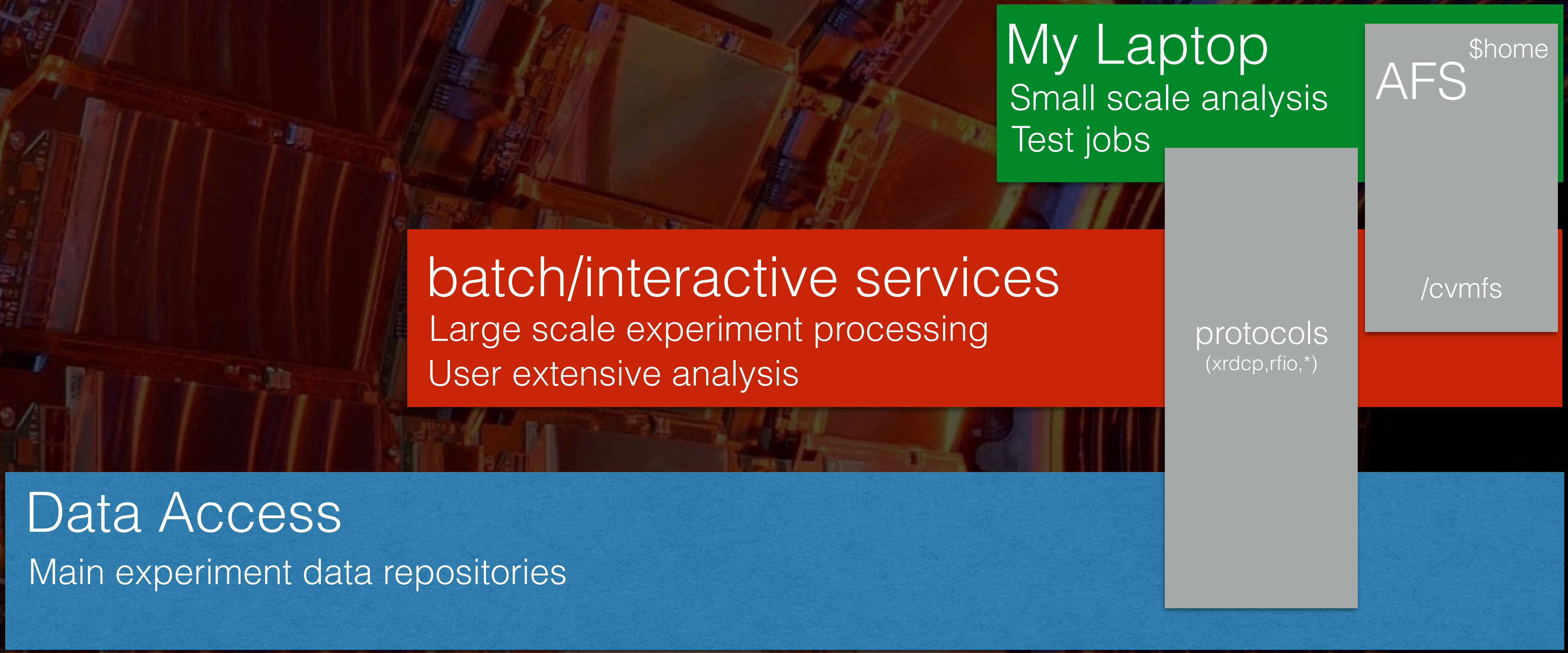
**+30% / yr tape capacity per $**

**+20%/yr I/O increase**

IT-ST

# Goals

## Make data access easy



My Laptop

Small scale analysis

Test jobs

AFS

$home

/cvmfs

batch/interactive services

Large scale experiment processing

User extensive analysis

protocols

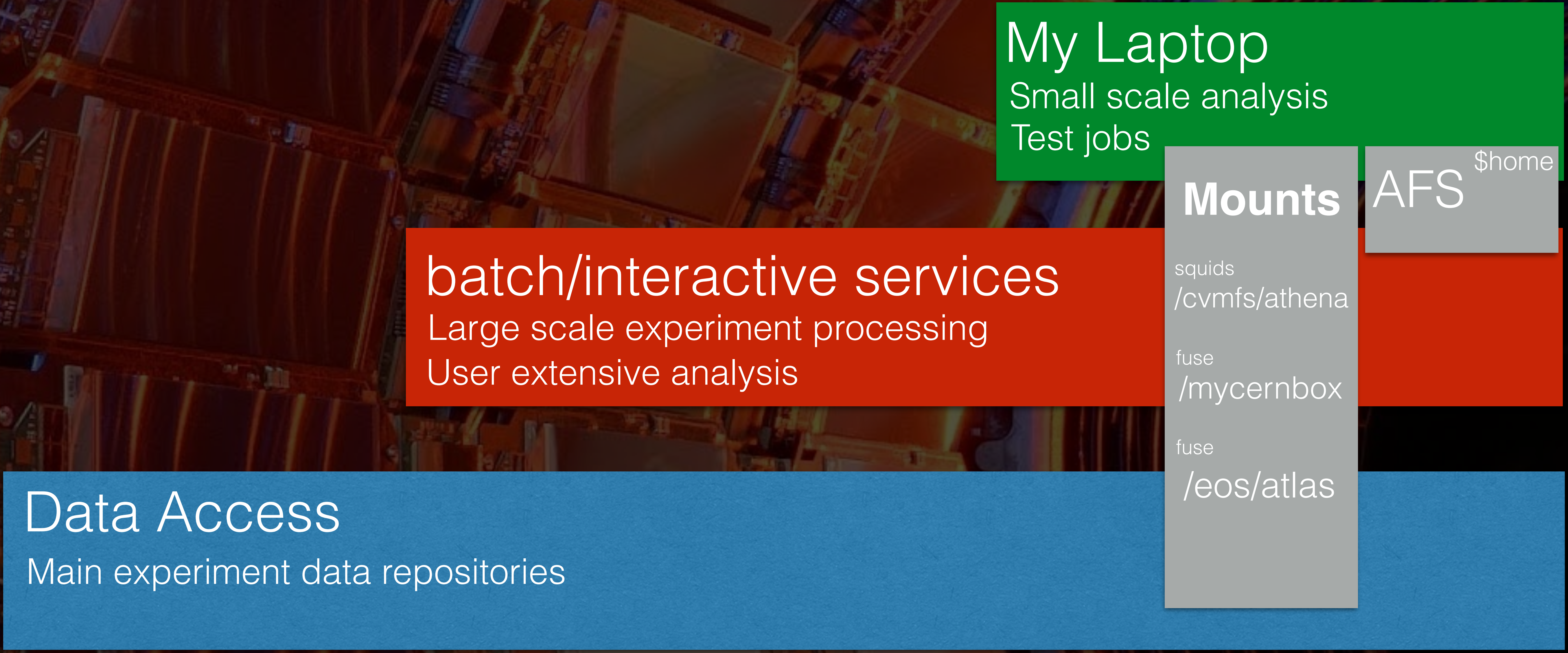
(xrdcp,rfio,*)

Data Access

Main experiment data repositories

# Goals

## Make data access easy



My Laptop

Small scale analysis

Test jobs

AFS $home

batch/interactive services

Large scale experiment processing

User extensive analysis

Mounts

squids /cvmfs/athena

fuse /mycernbox

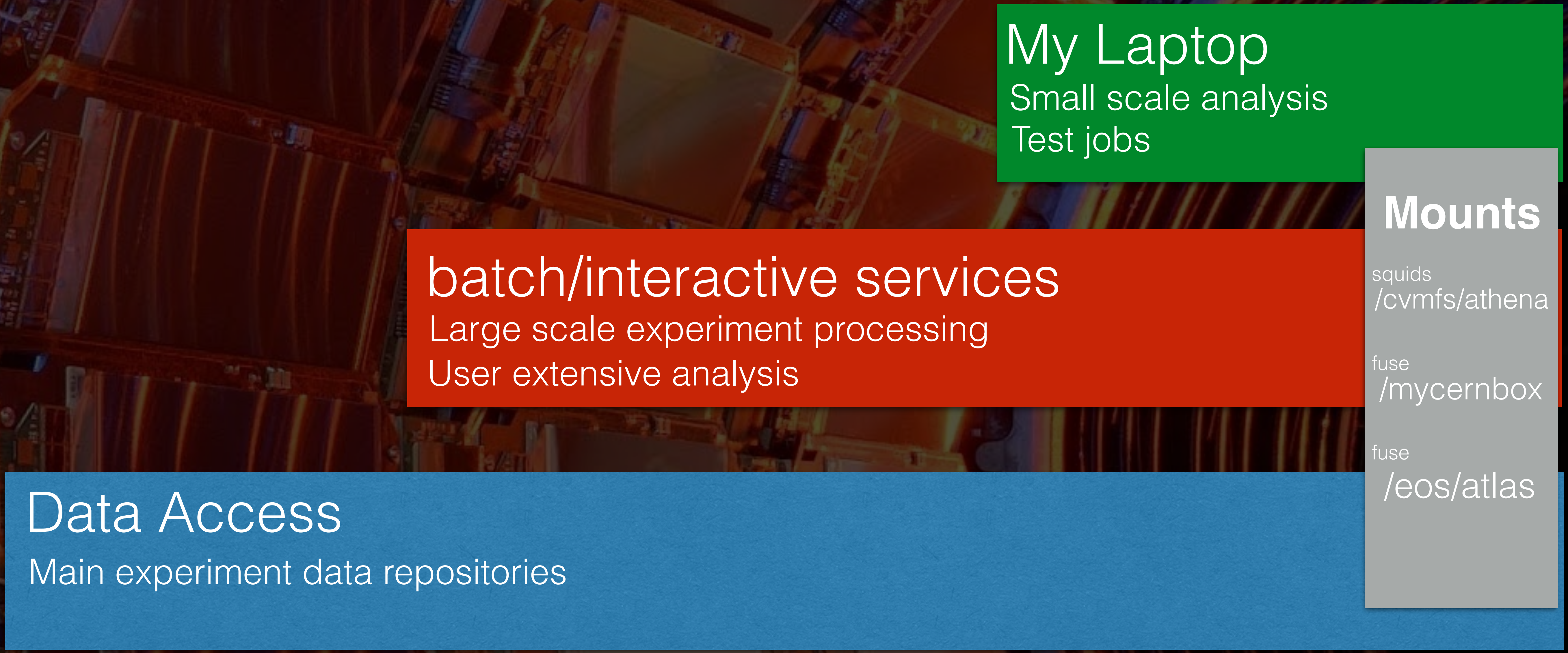
fuse /eos/atlas

Data Access

Main experiment data repositories

Goals

# Make data access easy



My Laptop

Small scale analysis

Test jobs

batch/interactive services

Large scale experiment processing

User extensive analysis

Data Access

Main experiment data repositories

**Mounts**

squids /cvmfs/athena

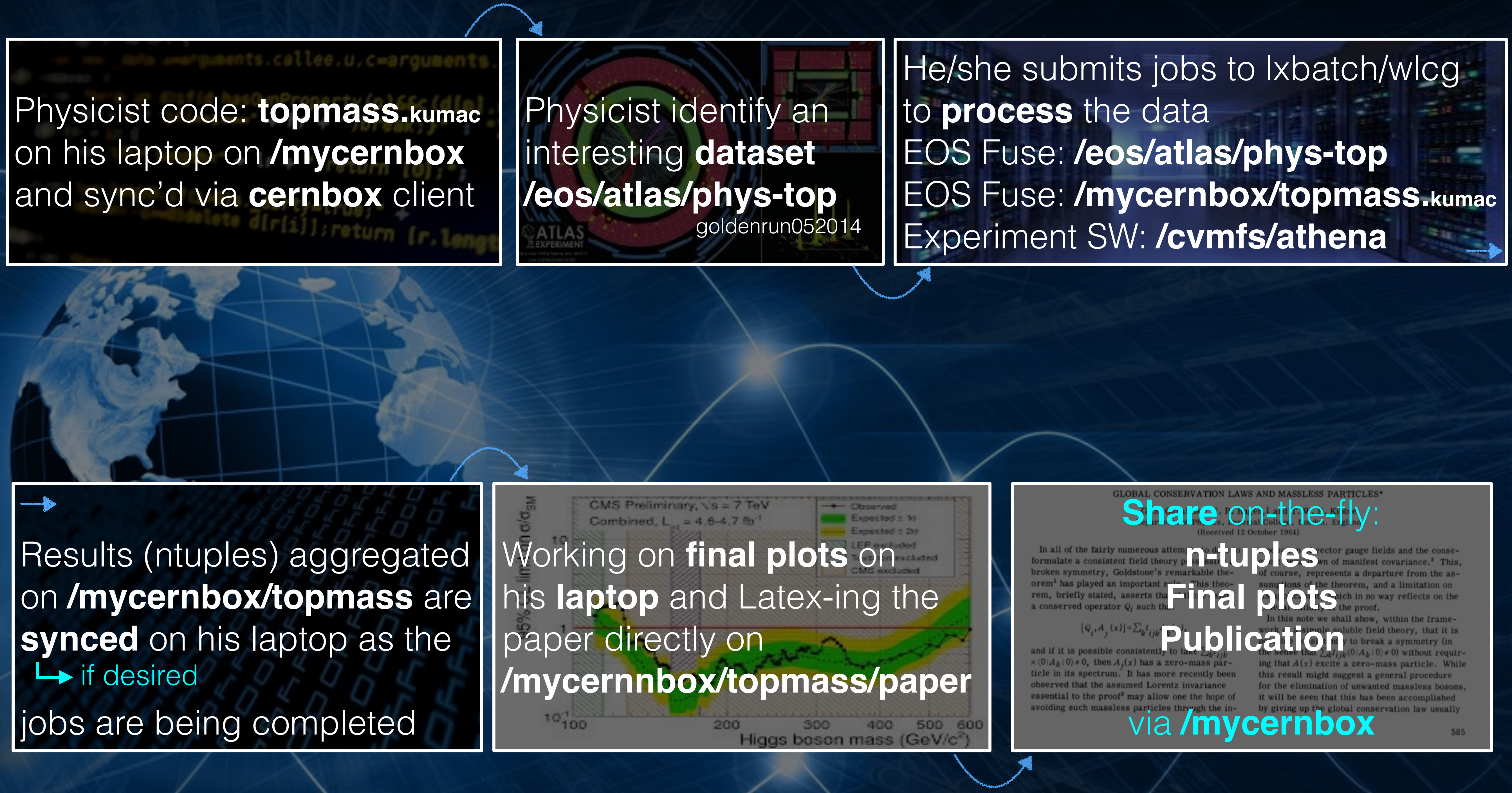
fuse /mycernbox

fuse /eos/atlas

EOS CERNBOX does *“your files”* /cernbox/jdoe
EOS ”*experiment”* does *“big data”* /eos/lhcb
Different QoS, different patterns, overlaps

# Goals

# Make Analysis Simple



Physicist code: **topmass.kumac** on his laptop on **/mycernbox** and sync'd via **cernbox** client

Physicist identify an interesting **dataset** **/eos/atlas/phys-top** goldenrun052014

He/she submits jobs to lxbatch/wlcg to **process** the data
EOS Fuse: **/eos/atlas/phys-top**
EOS Fuse: **/mycernbox/topmass.kumac**
Experiment SW: **/cvmfs/athena**

Results (ntuples) aggregated on **/mycernbox/topmass** are **synced** on his laptop as the
↳ if desired
jobs are being completed

Working on **final plots** on his **laptop** and Latex-ing the paper directly on **/mycernnbox/topmass/paper**

**Share** on-the-fly:
**n-tuples**
**Final plots**
**Publication**
via **/mycernbox**

ATLAS EXPERIMENT
CMS Preliminary, √s = 7 TeV
Higgs boson mass (GeV/c²)
GLOBAL CONSERVATION LAWS AND MASSLESS PARTICLES*

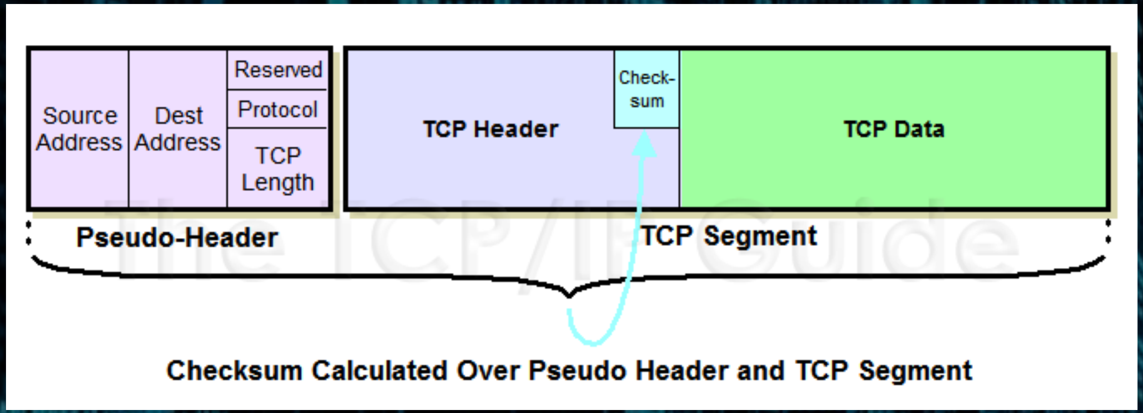
EOS is the enabling technology binding all this

Multi QoS   Access patterns   Protocols   Redundancy

CERN IT-ST

# Before concluding… D4T4 C0RRUP710N Adler32?

Found corrupted files after an incident caused by a a faulty router



Checksum Calculated Over Pseudo Header and TCP Segment

Brute force retransmits eventually went through due to TCP checksum collision

adler32 effectively gives less than 32-bits of verificaton power on file transfers affected by TCP checksoum collissions

(atlas) 7.8K files, 11 TB were staged in from T1 tapes for a md5 comparison, 17 corruptions with a colliding adler32 and a different md5 checksum —> 0.22% silent corruption rate

Observed courrption probability of 0.22% —> $2^{-9}$ (wrt. $2^{-32}$)

http://cern.ch/go/wr8j

# Goals
summary

Ensure a coherent development and operation of storage services at CERN for all aspects of physics data

Keep developing and operating Storage Services for Physics at the highest level

Communicating
Understanding
Delivering

Keep the ability to adapt and react fast

Problem/solution
Ask/Implement
In-house knowhow

Evaluate and investigate evolutions in technologies for better service/$

More for less
Operational costs
New applications

“Envision“ new models on data mananagement and analysis

LHC@myPC
Sync&Share
DmaaS

CERN