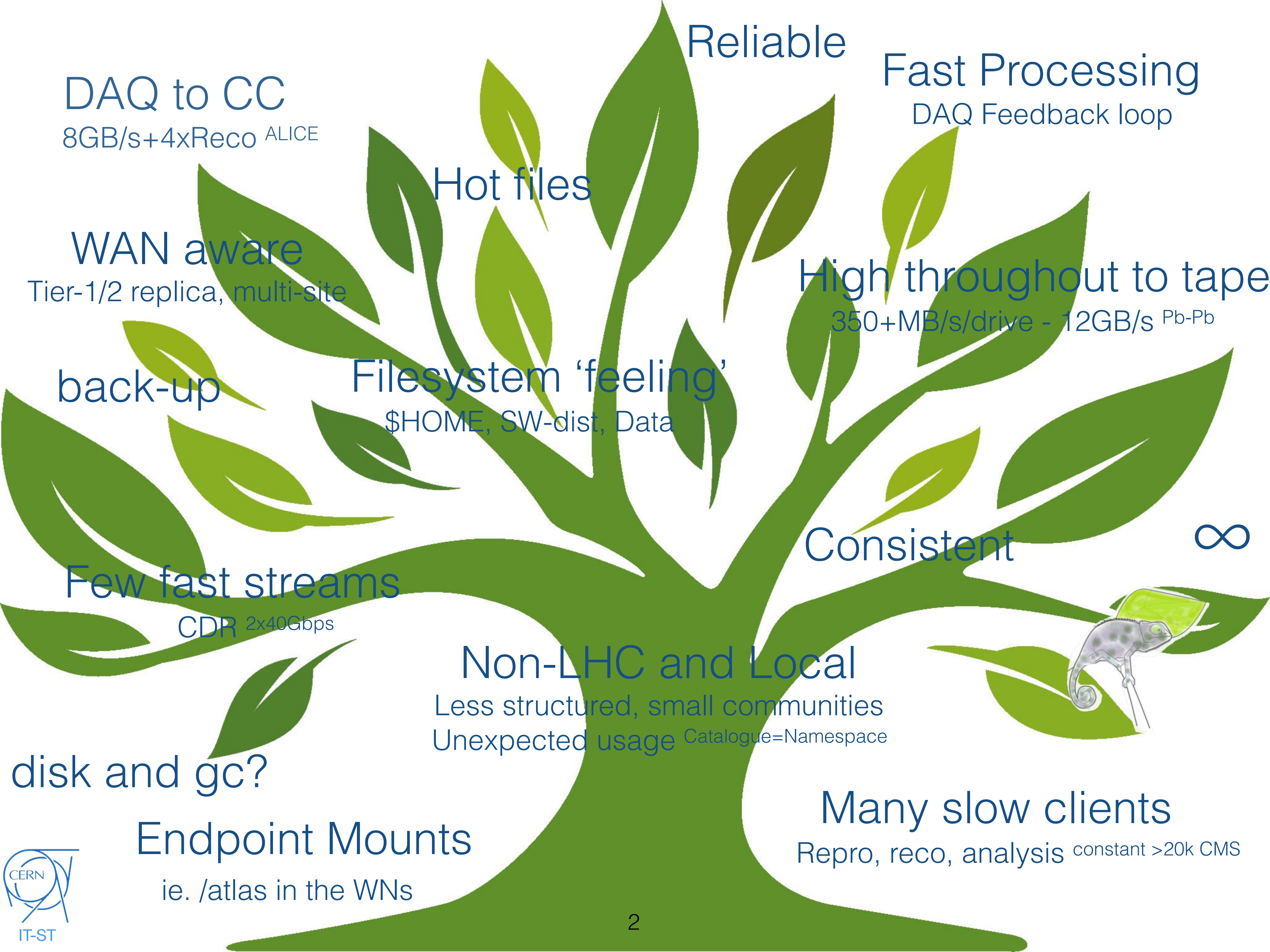
CERN/IT Storage Evolution

CERN

Xavier Espinal

on behalf of IT/ST



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

Evolved to

Tape oriented system

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

High throughput from DAQ, high throughput to tape

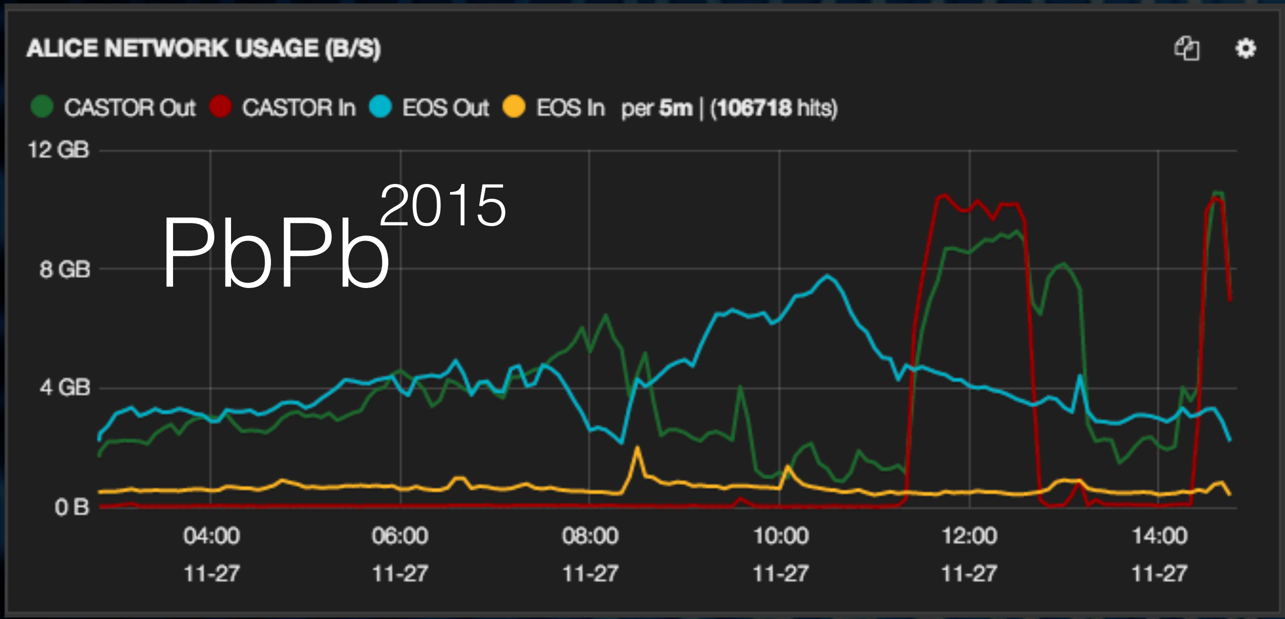
Key feature

Per stream speed

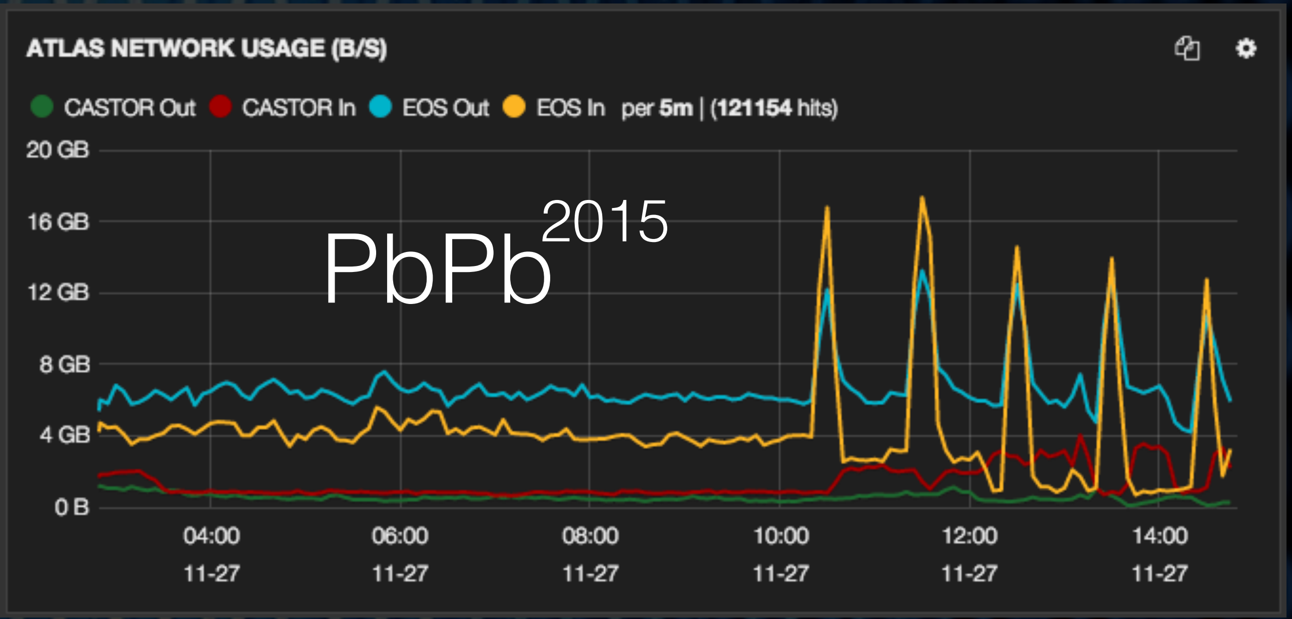
Moved from Raid1 to Raid60 (100MB/s to >350MB/s$^{\text{per stream}}$)

Evaluating common disk layer

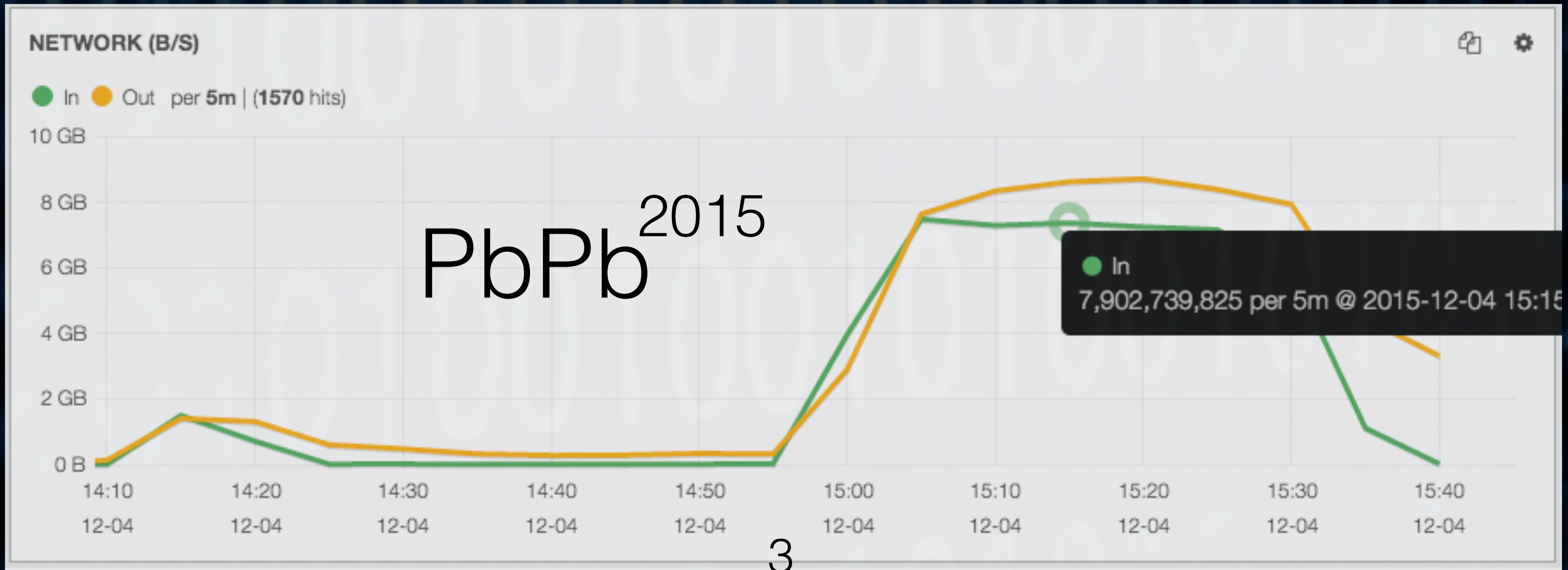
Tape policies, per experiment/user/group resources



PbPb^{2015}



PbPb^{2015}



PbPb^{2015}

CERN
IT-ST

CASTOR
CERN Advanced STORage manager

Evolved to

Tape oriented system

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

High throughput from DAQ, high throughput to tape

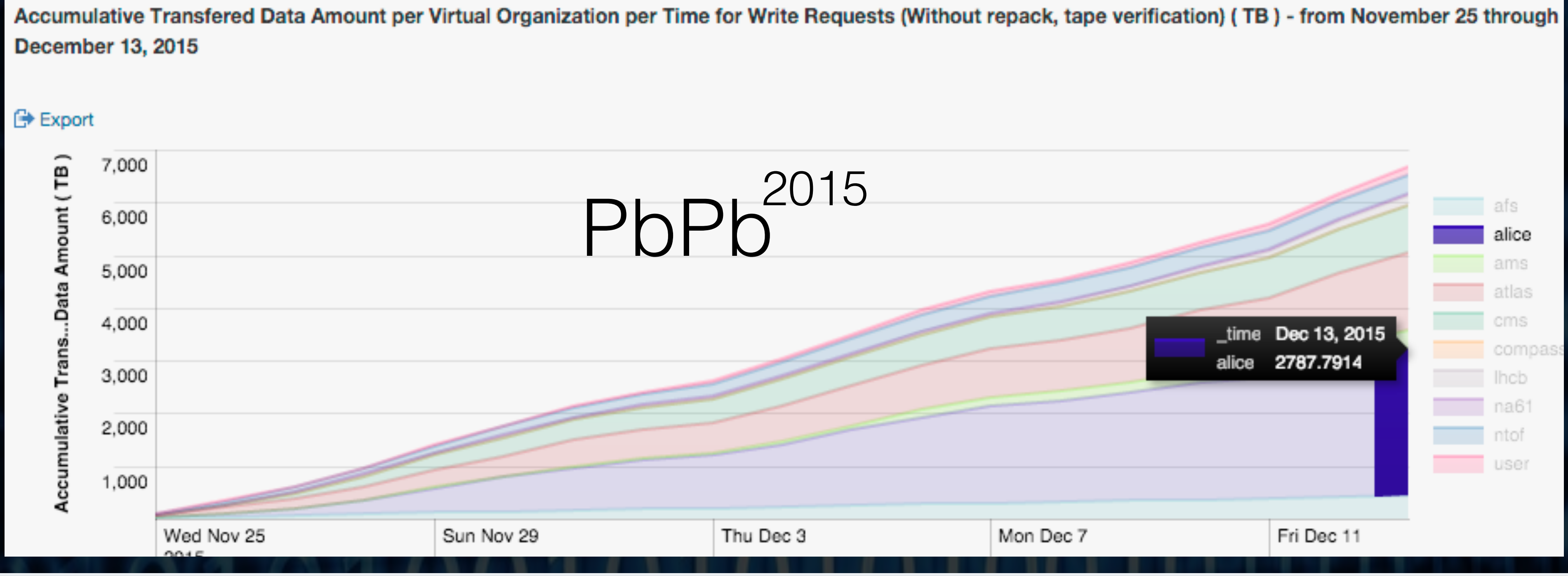
Key feature

Per stream speed

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

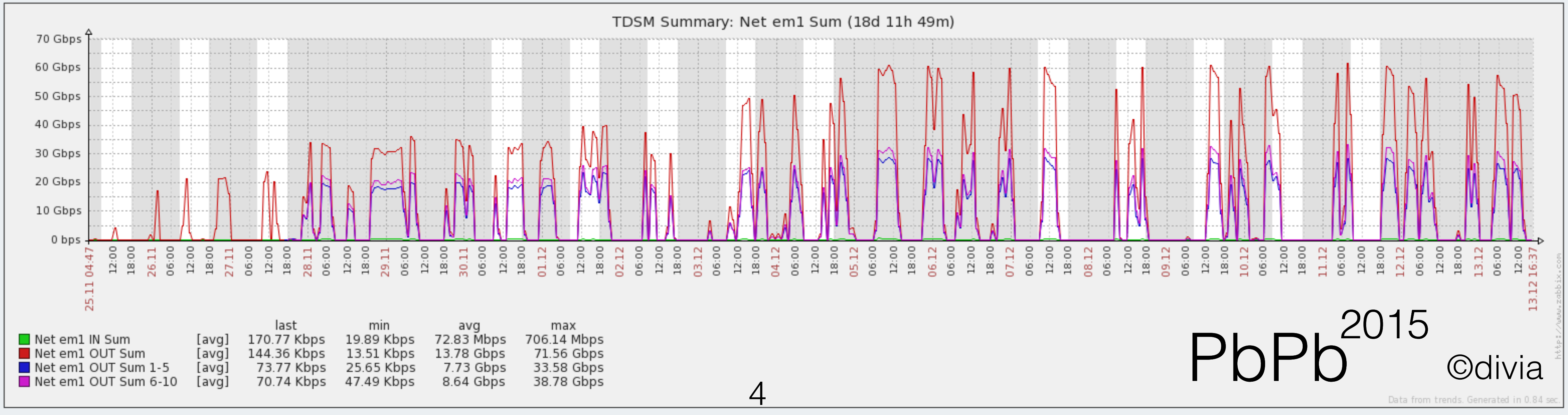
Evaluating common disk layer

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

PbPb^{2015}

©divia

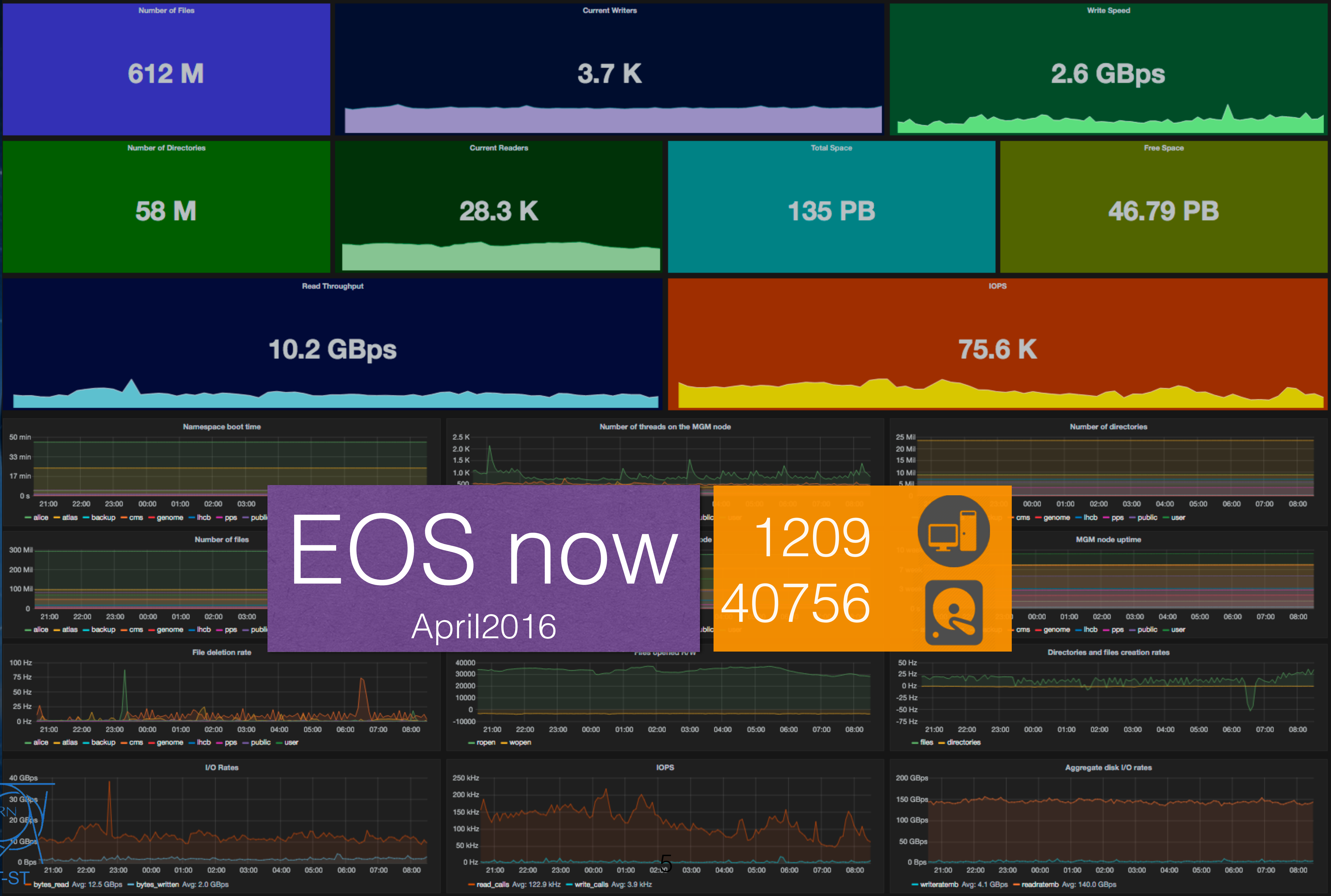
CERN
IT-ST

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

Namespace boot time
Number of threads on the MGM node
Number of directories
Number of files
MGM node uptime
File deletion rate
Files opened R/W
Directories and files creation rates
I/O Rates
IOPS
Aggregate disk I/O rates
alice atlas backup cms genome lhcb pps public user
ropen wopen
files directories
bytes_read Avg: 12.5 GBps bytes_written Avg: 2.0 GBps
read_calls Avg: 122.9 kHz write_calls Avg: 3.9 kHz
writeratemb Avg: 4.1 GBps readratemb Avg: 140.0 GBps
21:00 22:00 23:00 00:00 01:00 02:00 03:00 04:00 05:00 06:00 07:00 08:00

EOS now
April2016

1209
40756

made@CERN

Designed and tailored for experiments needs

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

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

Openlab/COMTRADE

Adaptable

Catering with different uses

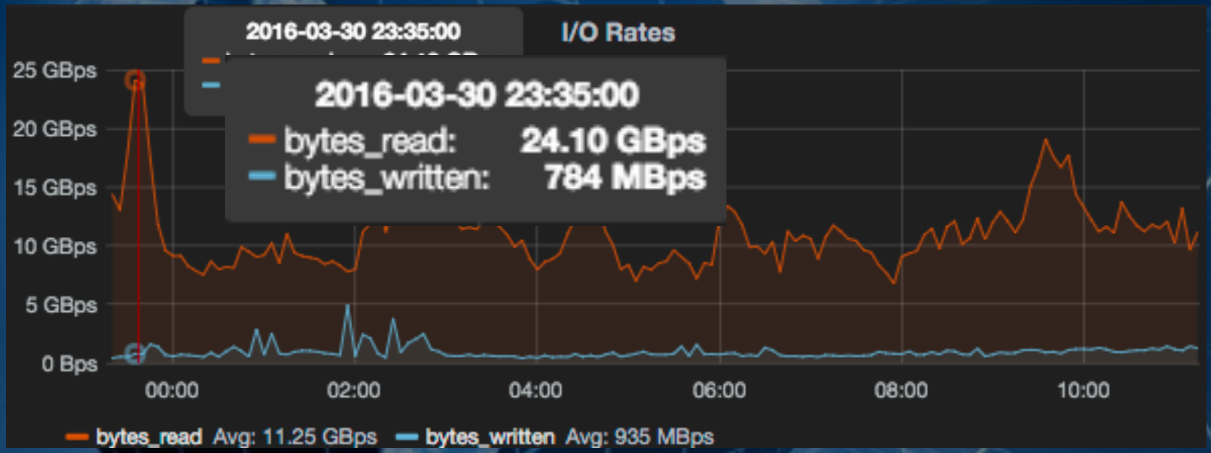
CDR
Data processing
User Analysis
CernBOX

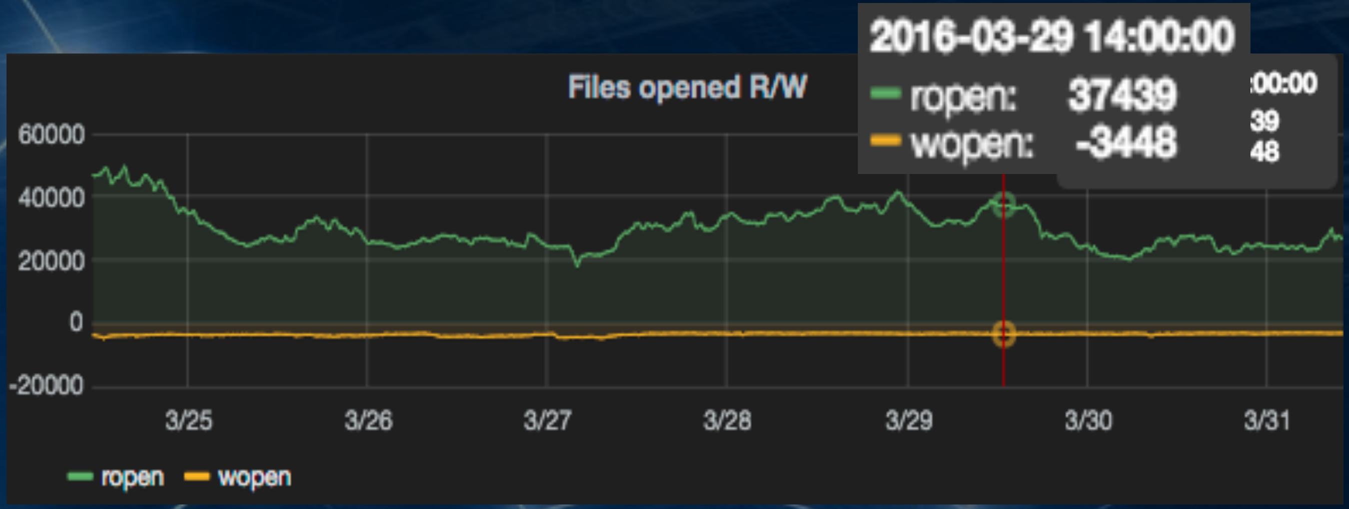
Community data

CERNBox

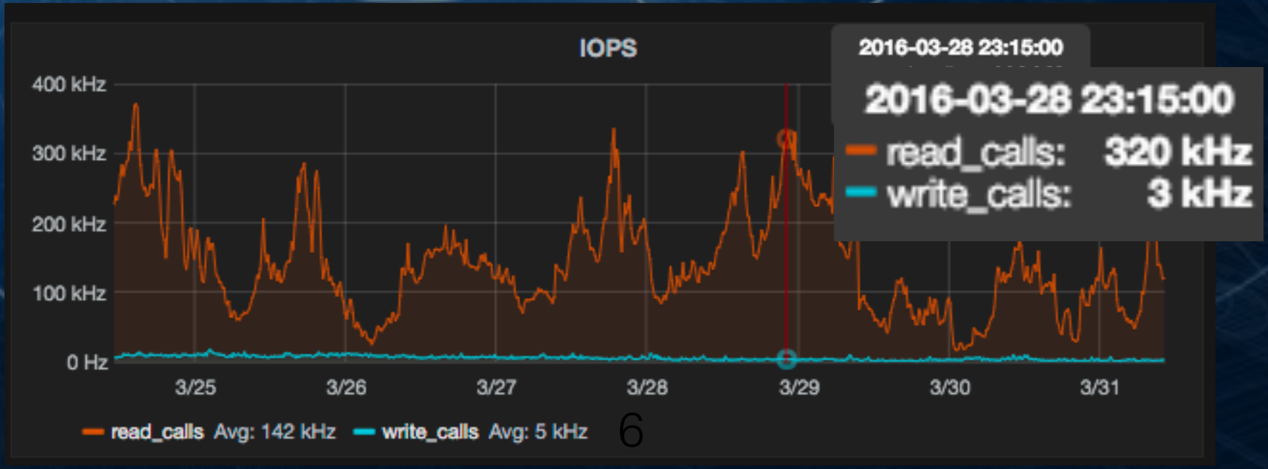
DmaaS (iJupyter)
Sync

share



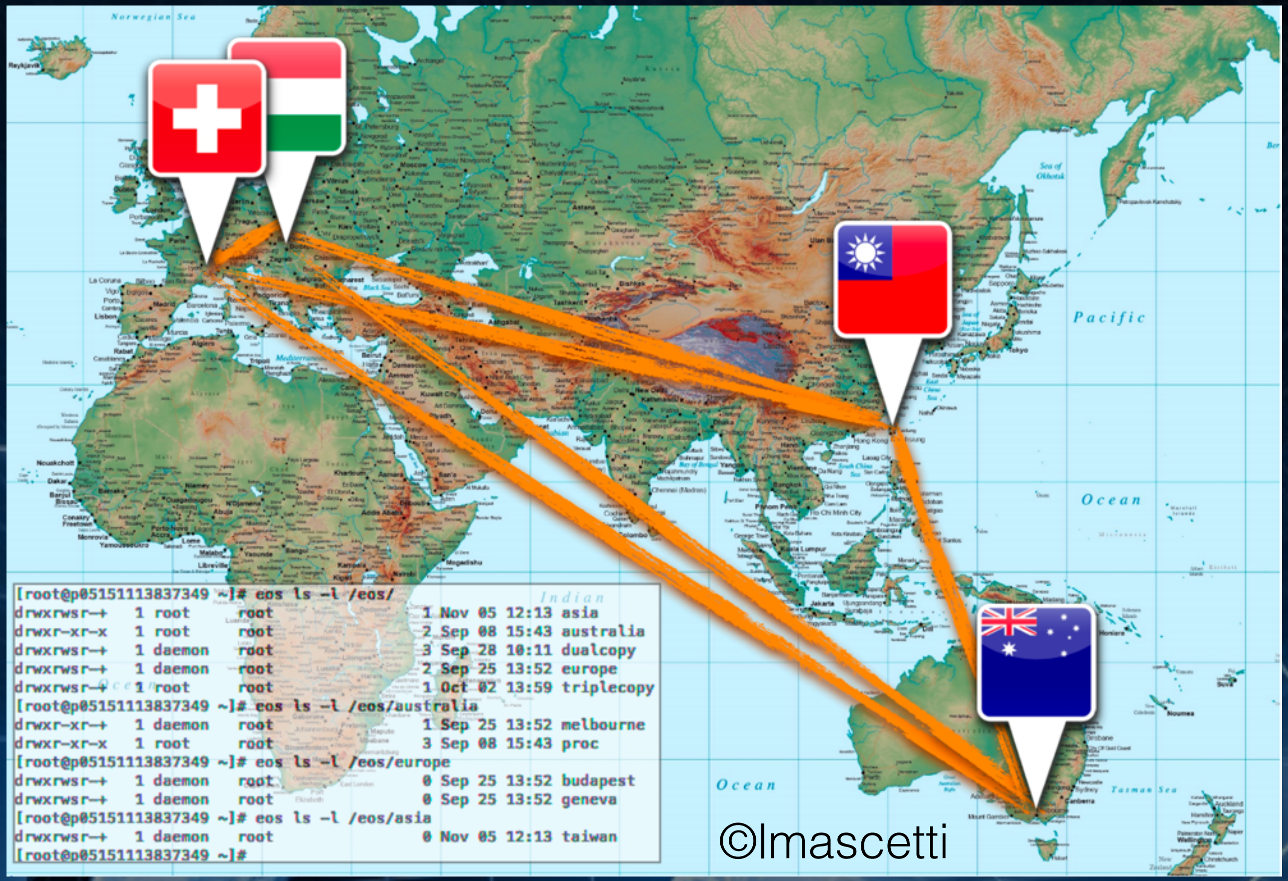


Future Shared FS ? more later…



©lmascetti

Can go distributed, can be shared and synced

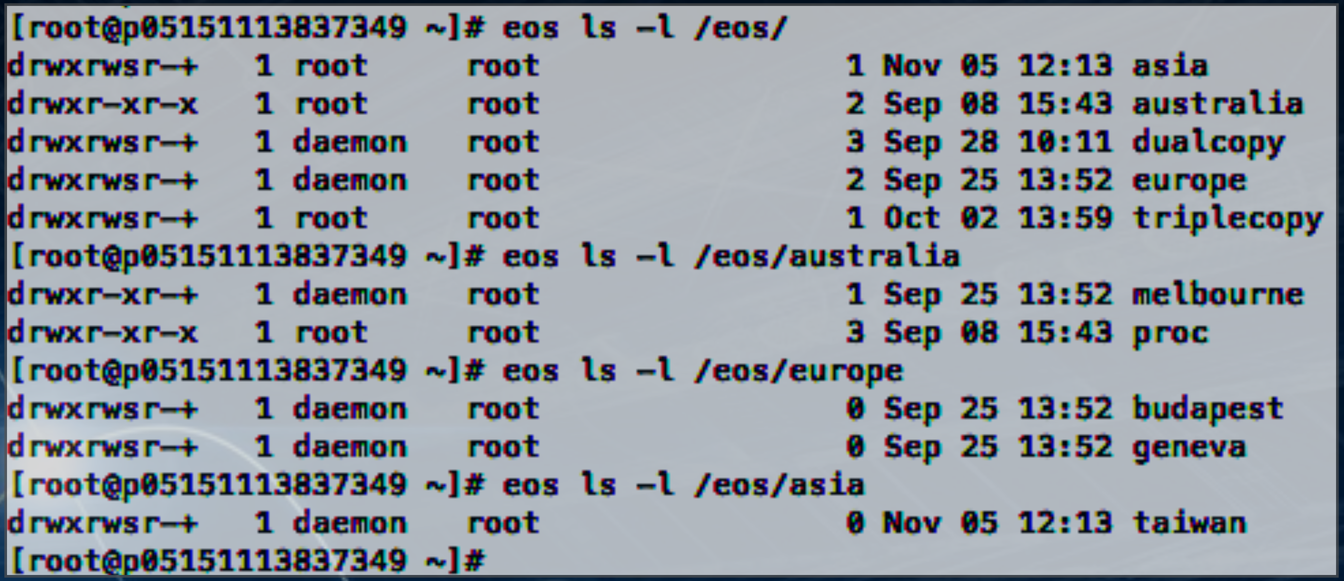

[root@p05151113837349 ~]# eos ls -l /eos/
drwxrwsr-+ 1 root root 1 Nov 05 12:13 asia
drwxr-xr-x 1 root root 2 Sep 08 15:43 australia
drwxrwsr-+ 1 daemon root 3 Sep 28 10:11 dualcopy
drwxrwsr-+ 1 daemon root 2 Sep 25 13:52 europe
drwxrwsr-+ 1 root root 1 Oct 02 13:59 triplecopy
[root@p05151113837349 ~]#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

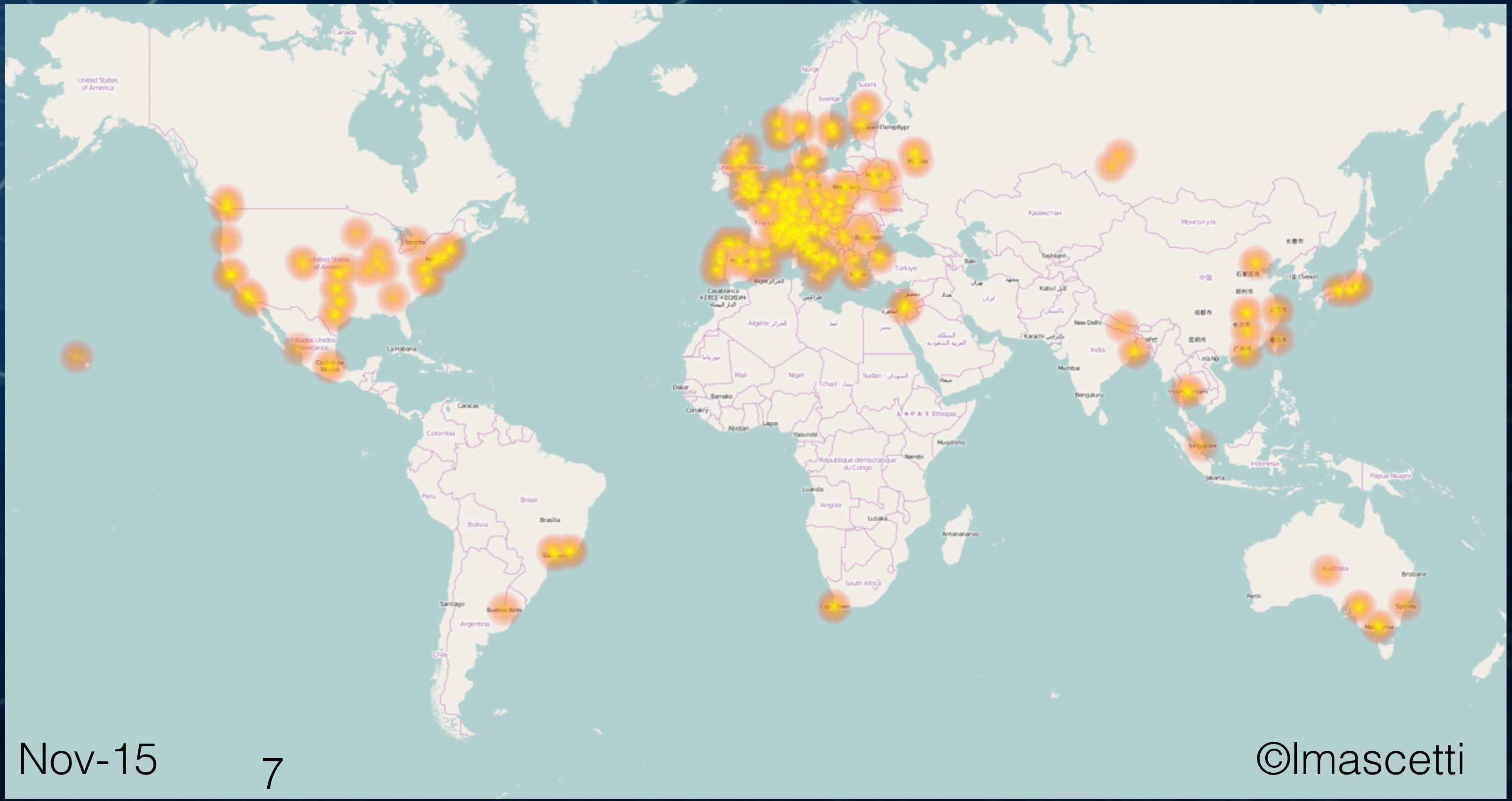
DmaaS (iJupyter)

Sync

CERNBox

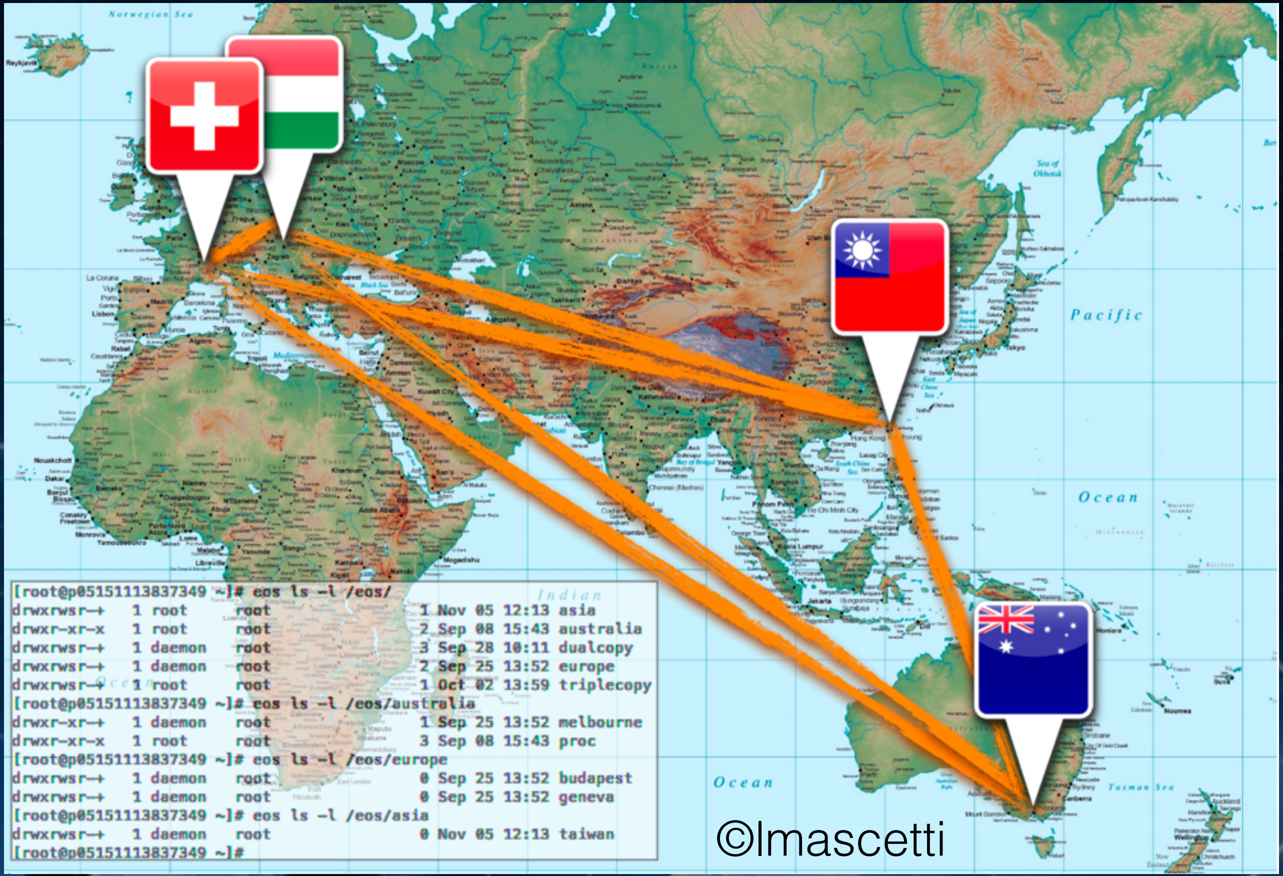
Users	4719
# files	70 Million
# dirs	9 Million
Quota	1TB/user
Used Space	125 TB
Deployed Space	1.5 PB

⊞ 20% 🍎 60% 🐧 20%



©lmascetti

Can go distributed, can be shared and synced



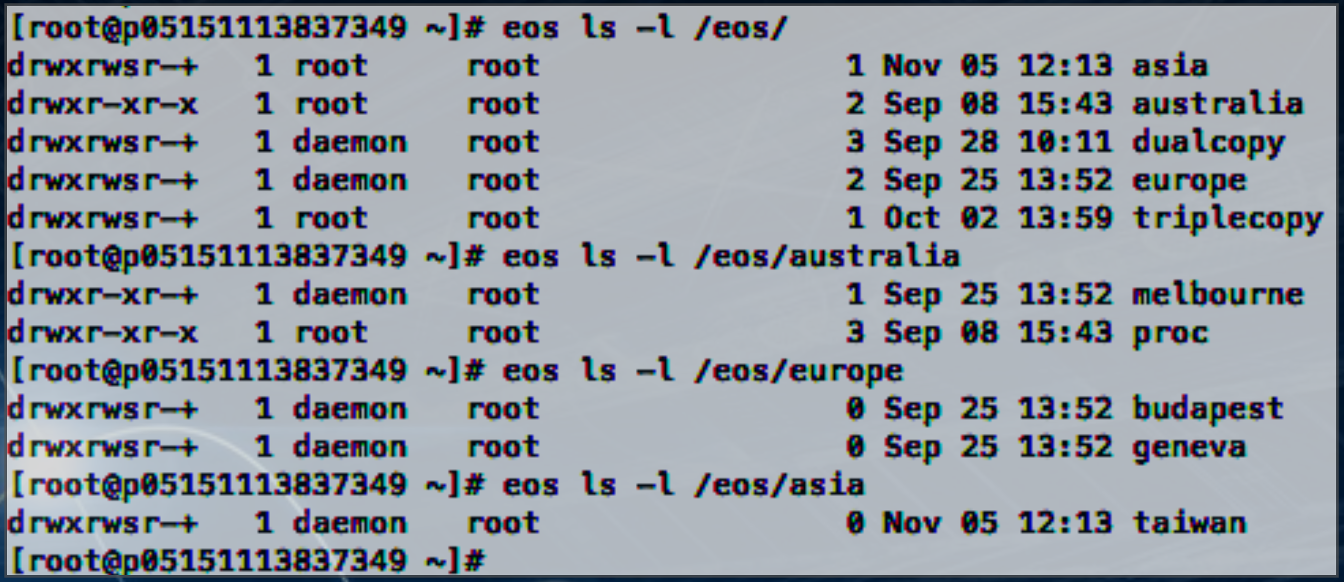
```
[root@p05151113837349 ~]# eos ls -l /eos/
drwxrwsr-+  1 root     root       1 Nov 05 12:13 asia
drwxr-xr-x  1 root     root       2 Sep 08 15:43 australia
drwxrwsr-+  1 daemon   root       3 Sep 28 10:11 dualcopy
drwxrwsr-+  1 daemon   root       2 Sep 25 13:52 europe
drwxrwsr-+  1 root     root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



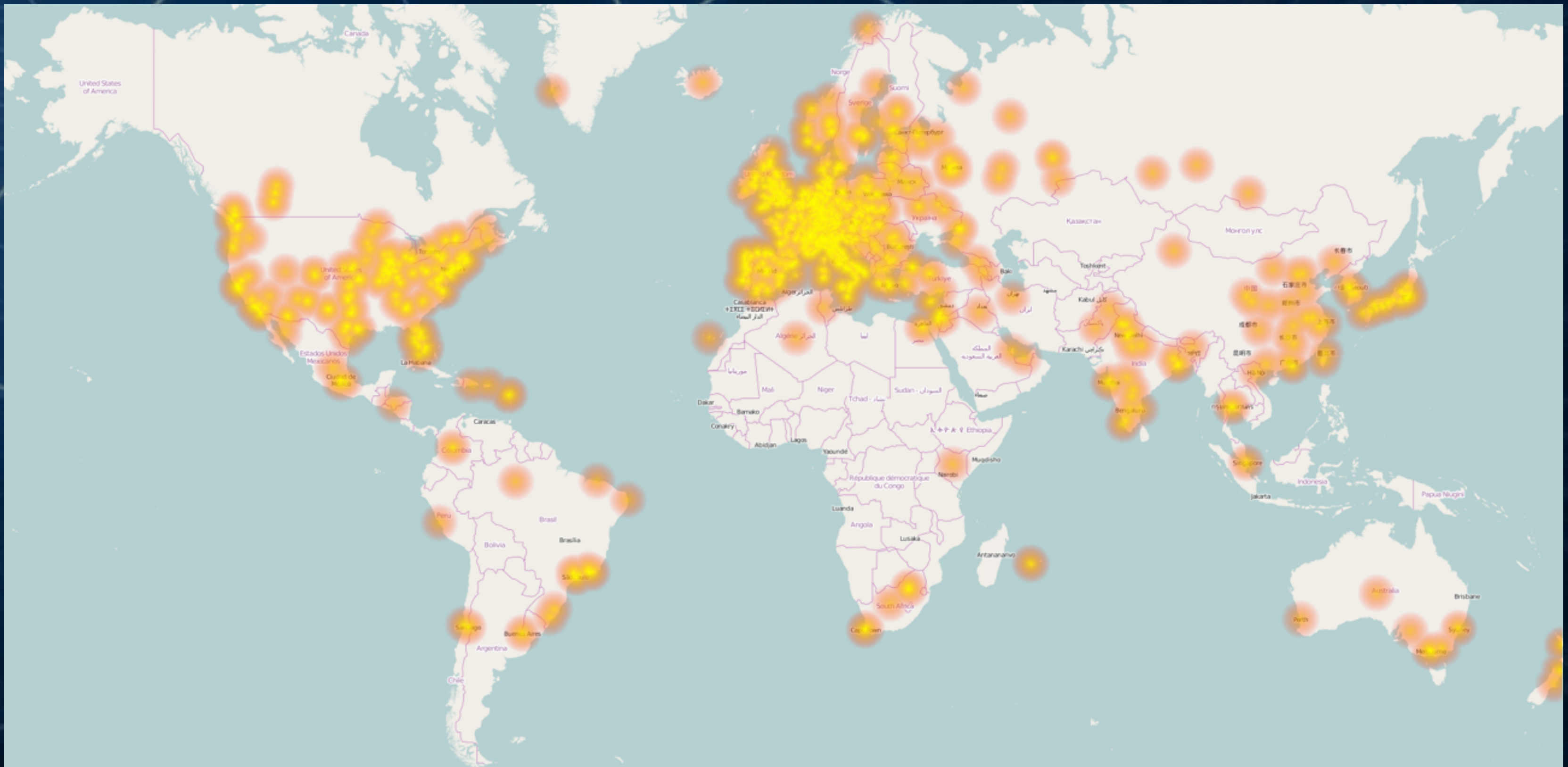
```
[root@p05151113837349 ~]# eos ls -l /eos/
drwxrwsr-+   1 root     root                  1 Nov 05 12:13 asia
drwxr-xr-x   1 root     root                  2 Sep 08 15:43 australia
drwxrwsr-+   1 daemon   root                  3 Sep 28 10:11 dualcopy
drwxrwsr-+   1 daemon   root                  2 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%

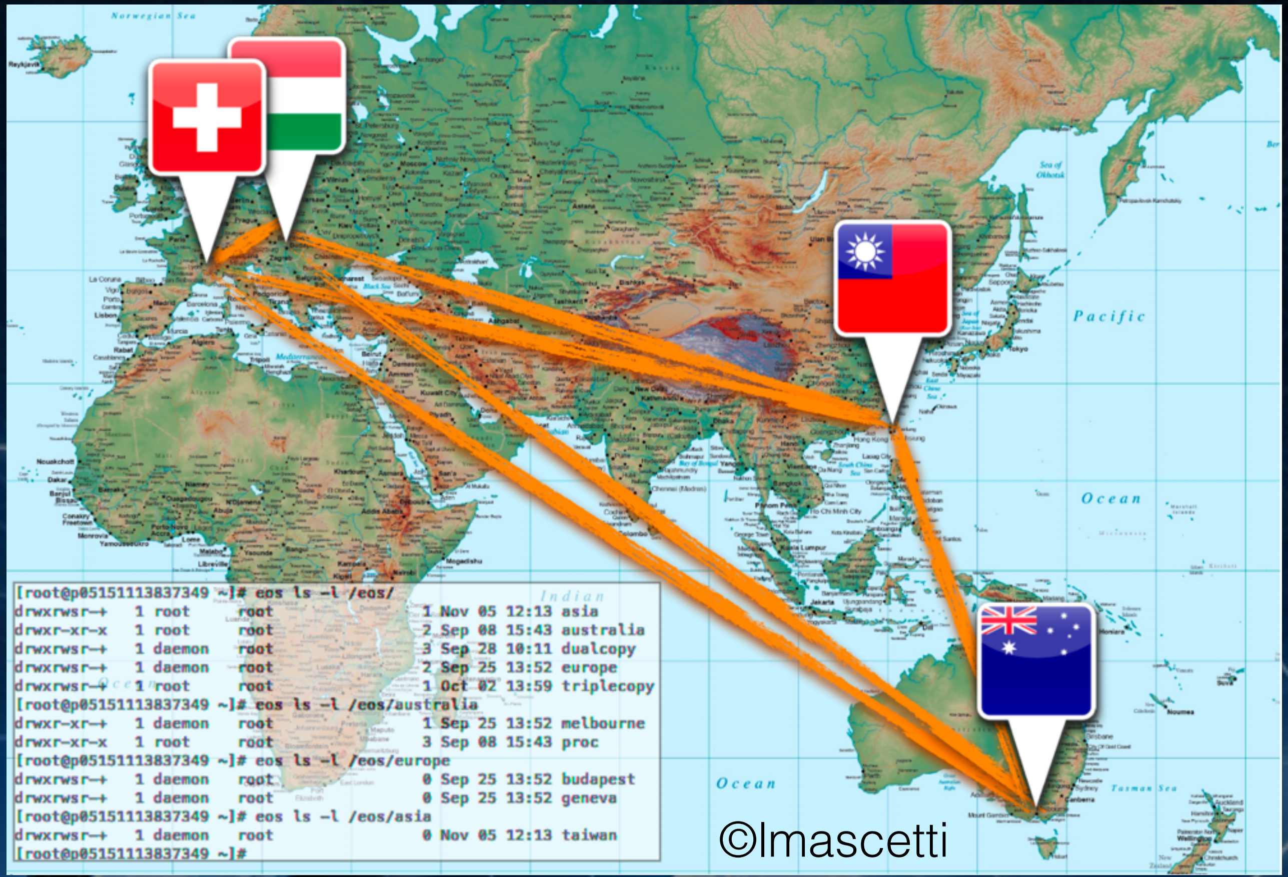


Dec-15

©lmascetti

EOS

Can go distributed, can be shared and synced



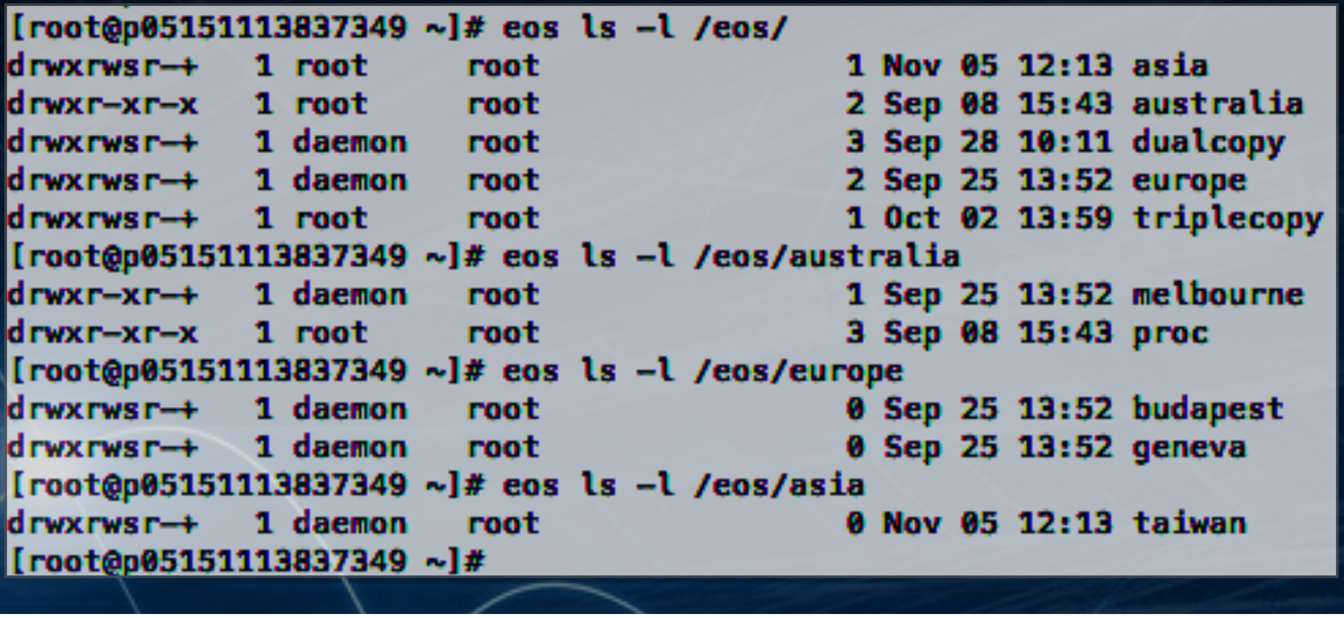
```
[root@p05151113837349 ~]# eos ls -l /eos/
drwxrwsr-+   1 root     root        1 Nov 05 12:13 asia
drwxr-xr-x   1 root     root        2 Sep 08 15:43 australia
drwxrwsr-+   1 daemon   root        3 Sep 28 10:11 dualcopy
drwxrwsr-+   1 daemon   root        2 Sep 25 13:52 europe
drwxrwsr-+   1 root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

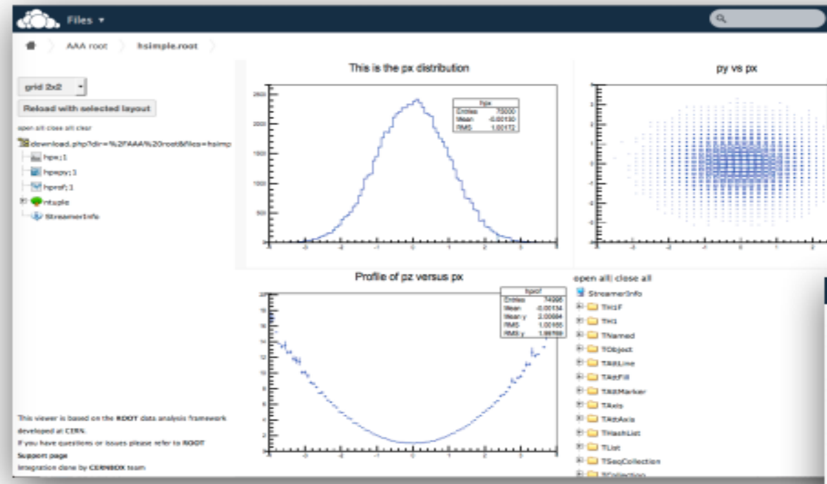
CERNBox

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

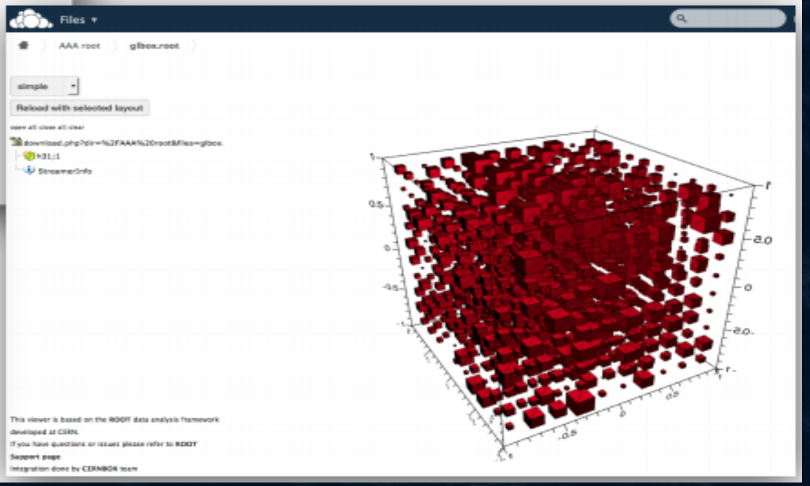
⊞ 20% 60% 🐧 20%

Embedded ROOT Viewer

©dpiparo



ROOT
Data Analysis Framework



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

Integration done by CERNBox team.

BLOCK STORAGE

Openstack VM
Cinder Volumes
S3

RADOS FS

File stripper
CASTOR backend
Under evaluation

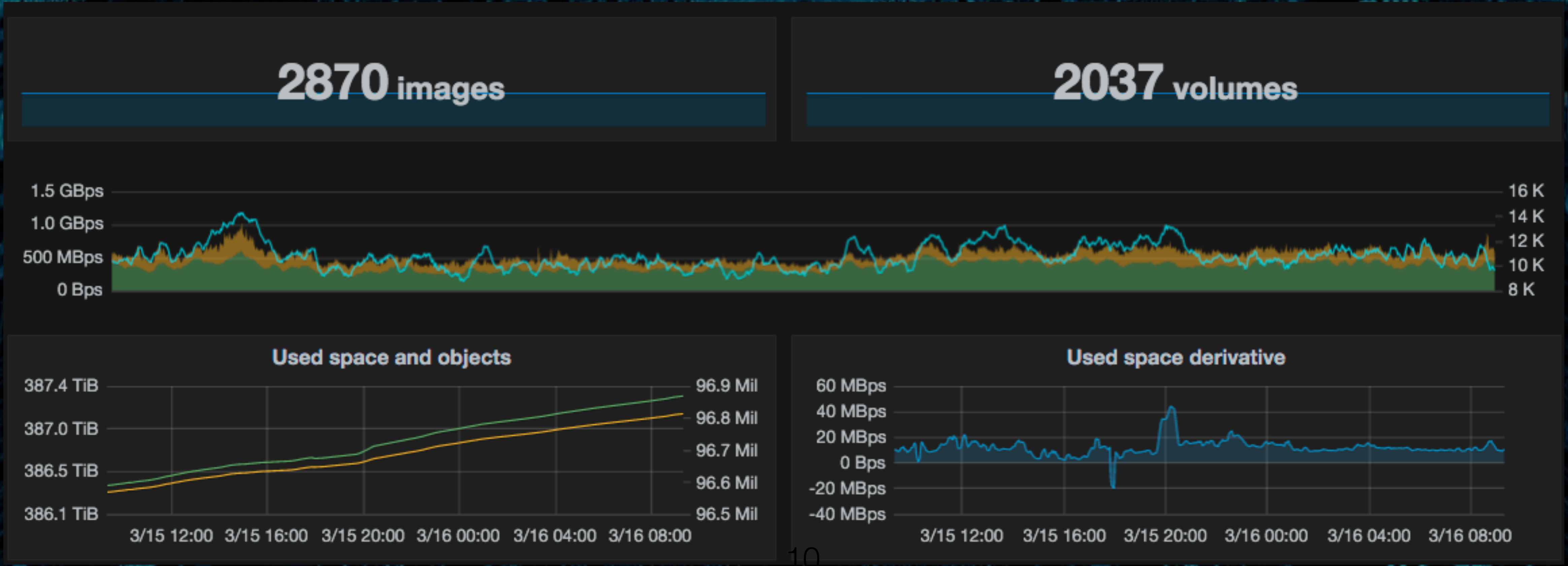
Large contribution
Community

Code
development
CERN-IT/ST

Largest Cluster 30PB
Deployed to date
40k OSDs

Multi-site
In production

3PB@wigner
1PB@meyrin



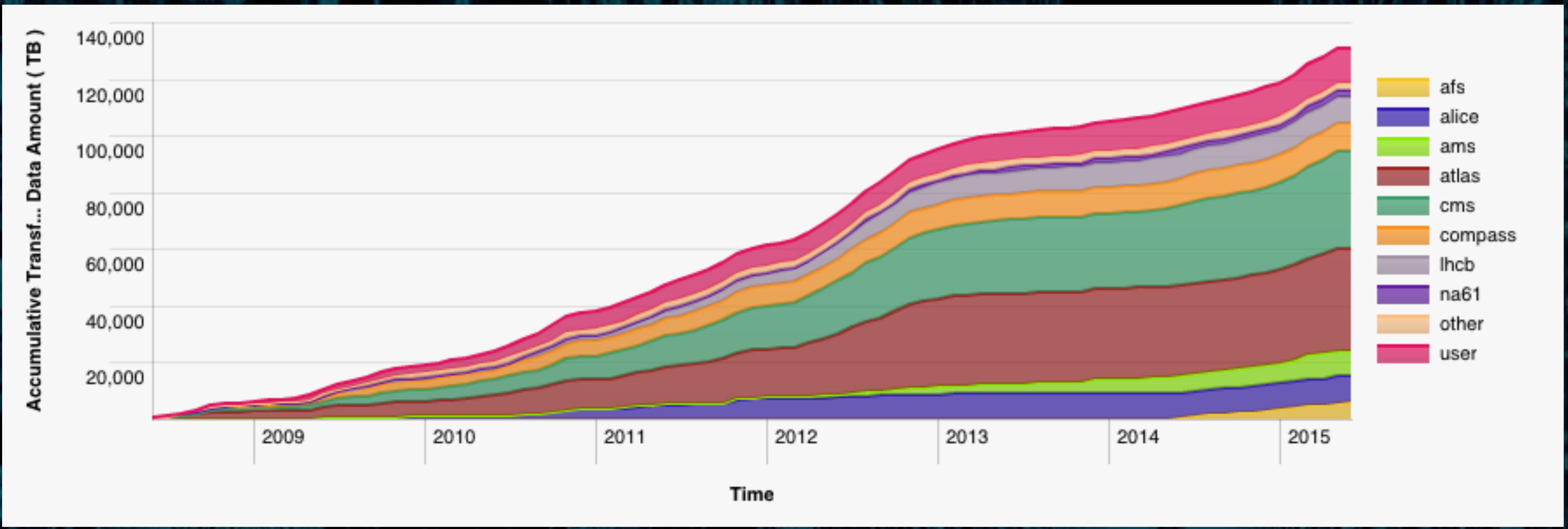
IT-ST

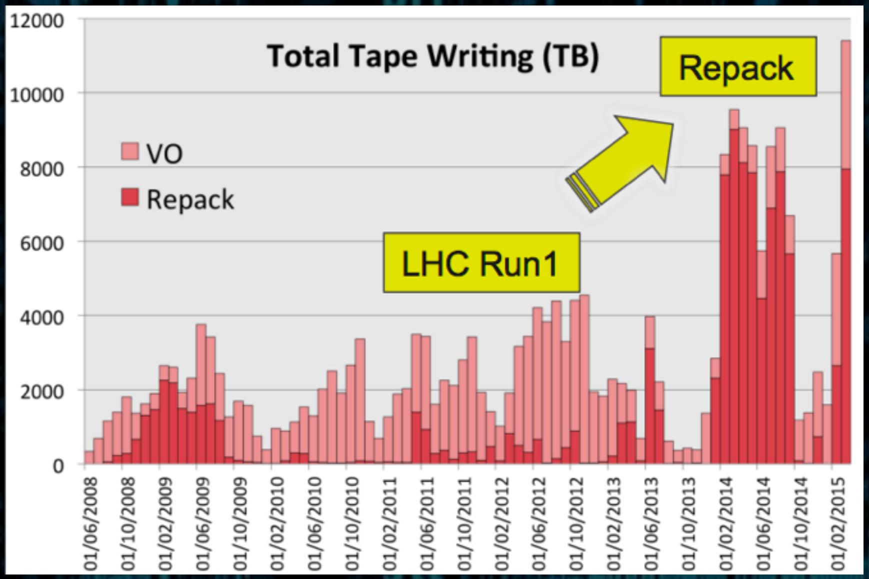
CERN Tape Archive

Technology driven: new medias brings ↑ density ↑ speed

Towards a pluggable tape backend (EOS)

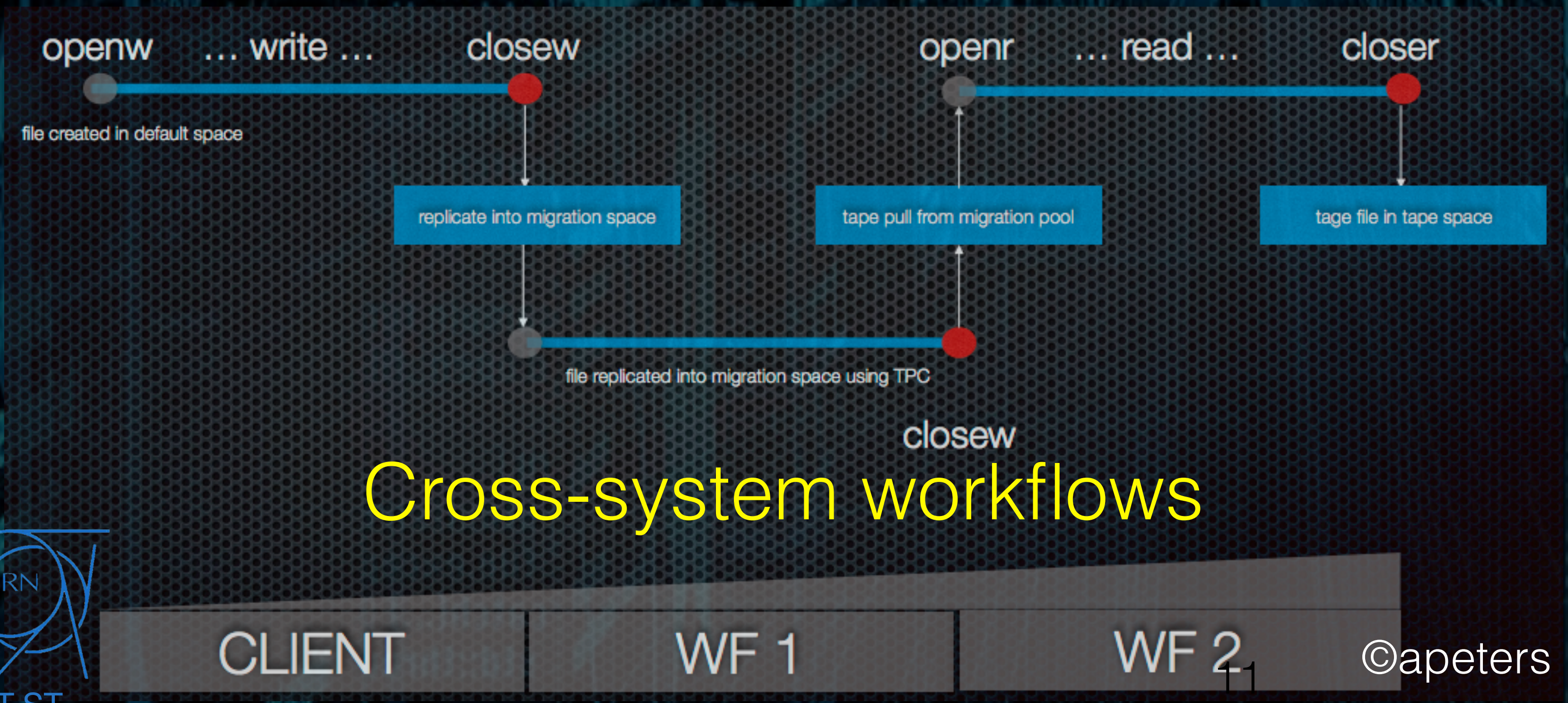
Cold by definition: hight throughput, high latency





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

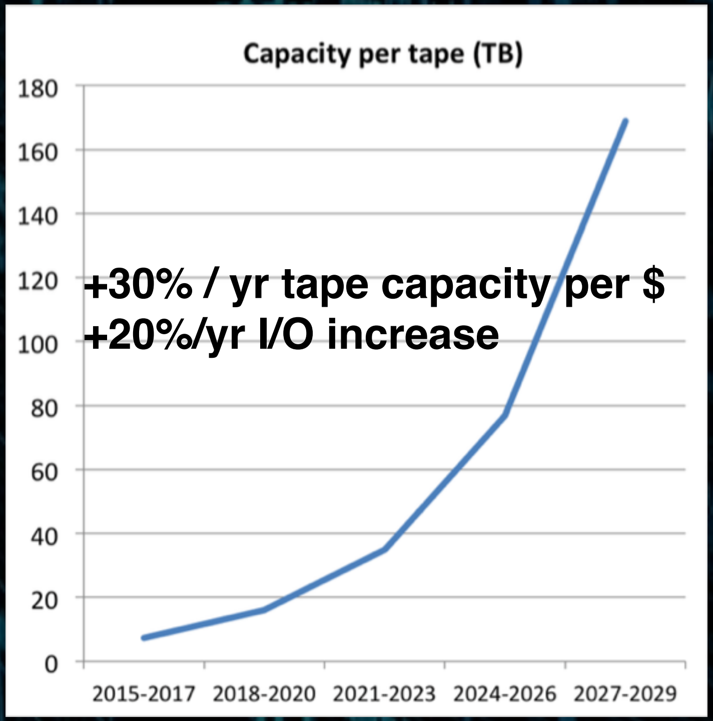
very large disk caches nowadays



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

Cross-system workflows

©apeters



+30% / yr tape capacity per $
+20%/yr I/O increase

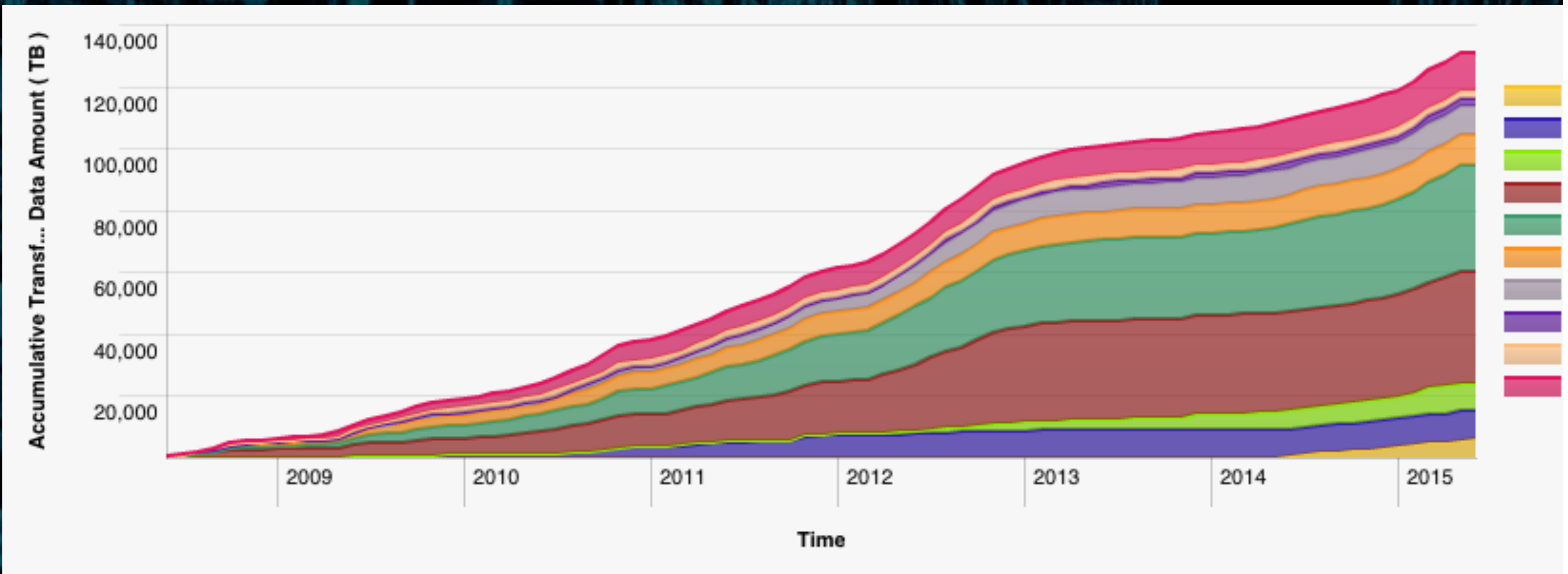
IT-ST

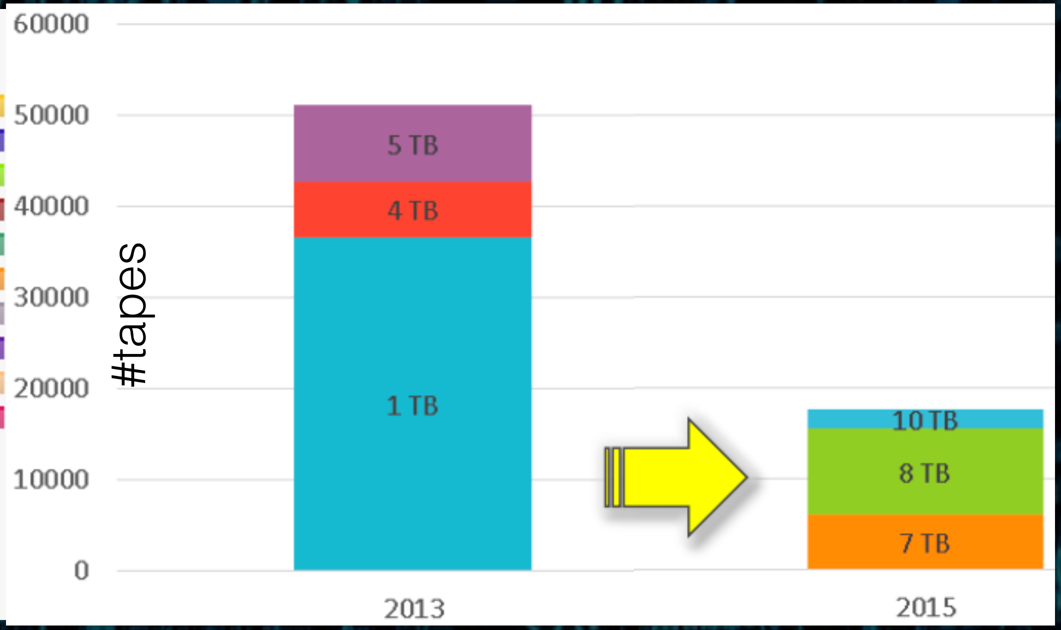
CERN Tape Archive

Technology driven: new medias brings ↑ density ↑ speed

Towards a pluggable tape backend (EOS)

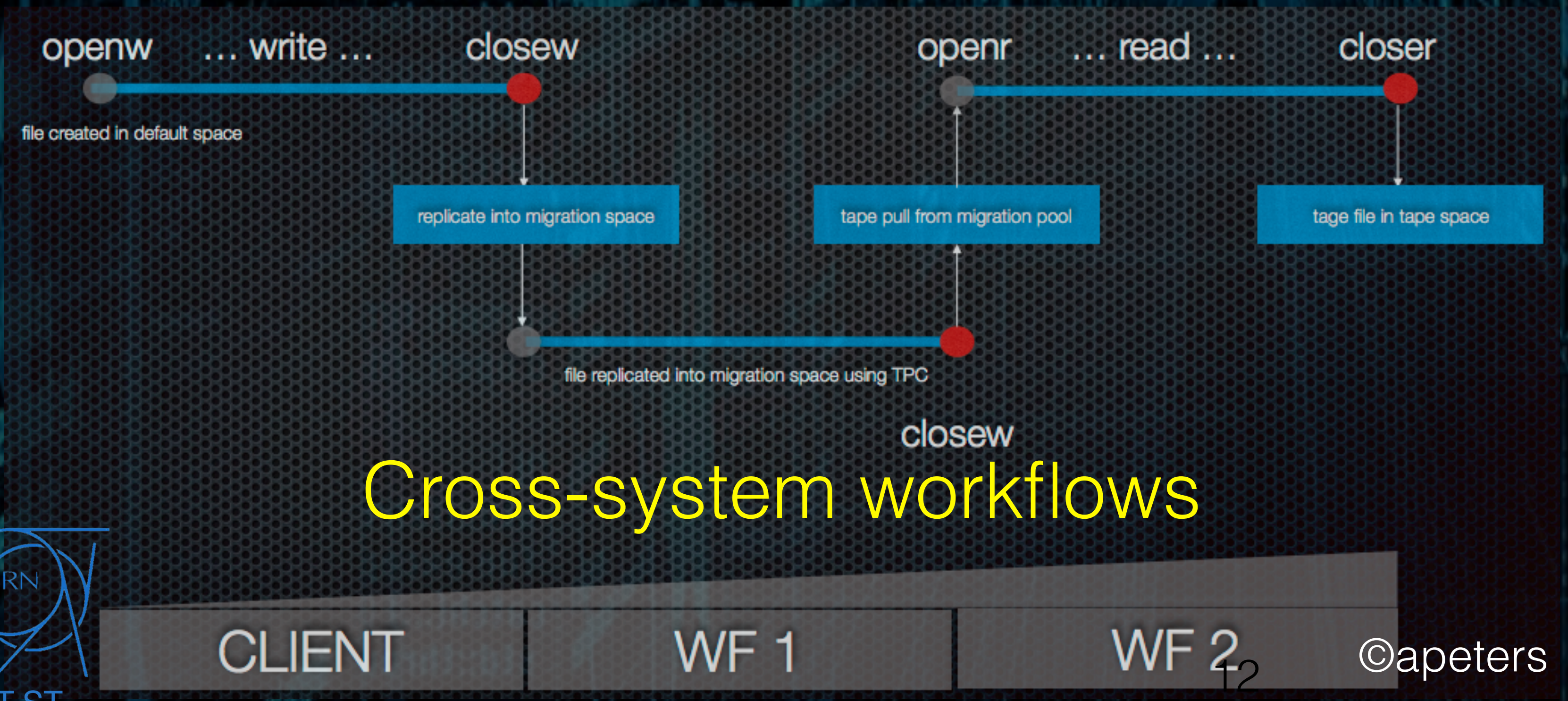
Cold by definition: hight throughput, high latency





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

very large disk caches nowadays



file created in default space

replicate into migration space

tape pull from migration pool

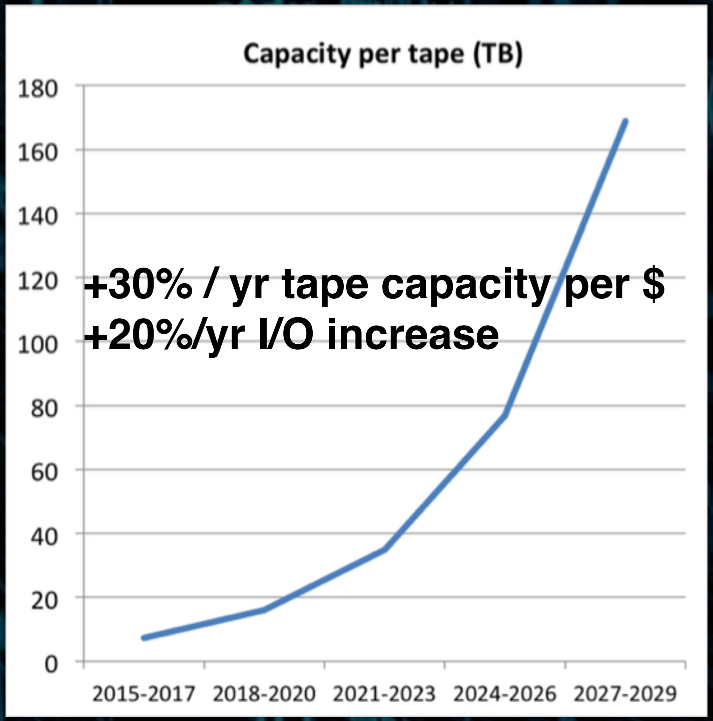
tage file in tape space

file replicated into migration space using TPC

closew

Cross-system workflows

©apeters



+30% / yr tape capacity per $
+20%/yr I/O increase

IT-ST

Goals

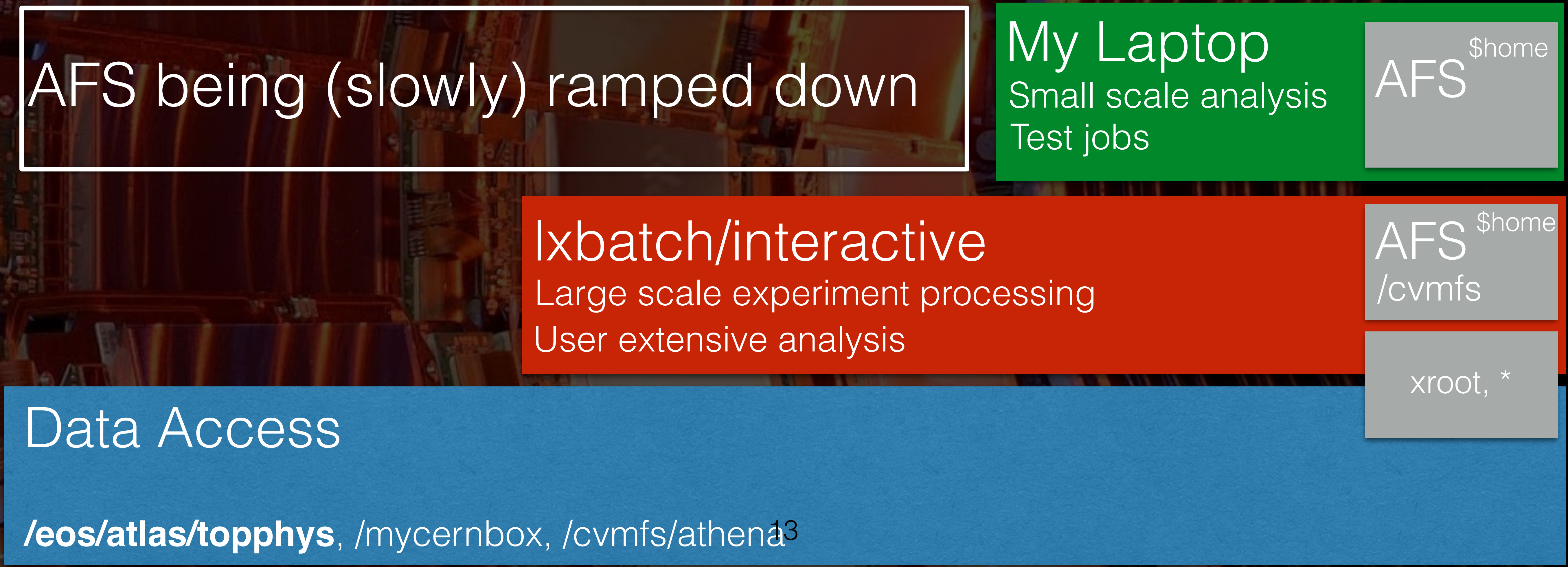
[1]Make data access easy

$HOME Batch /home
Laptop use →

DATA ACCESS Protocol based
xroot, rfio, etc. →

?

AFS being (slowly) ramped down



My Laptop
Small scale analysis
Test jobs

lxbatch/interactive
Large scale experiment processing
User extensive analysis

Data Access

/eos/atlas/topphys, /mycernbox, /cvmfs/athena

Goals

[1]Make data access easy

$HOME Batch /home
Laptop use →

DATA ACCESS Protocol based
xroot, rfio, etc. →

SharedFS /cernbox
Syncing option

Large Scale Storage Access MountPoints
/eos, /cernbox
/cvmfs

EOS CERNBOX does *"your files"* /cernbox/jdoe
EOS Experiment does *"big data"* /eos/lhcb
Different QoS, different patterns, overlaps
Backup

My Laptop

Small scale analysis
Test jobs

lxbatch/interactive

Large scale experiment processing
User extensive analysis

Mounts

/cvmfs/athena

/mycernbox

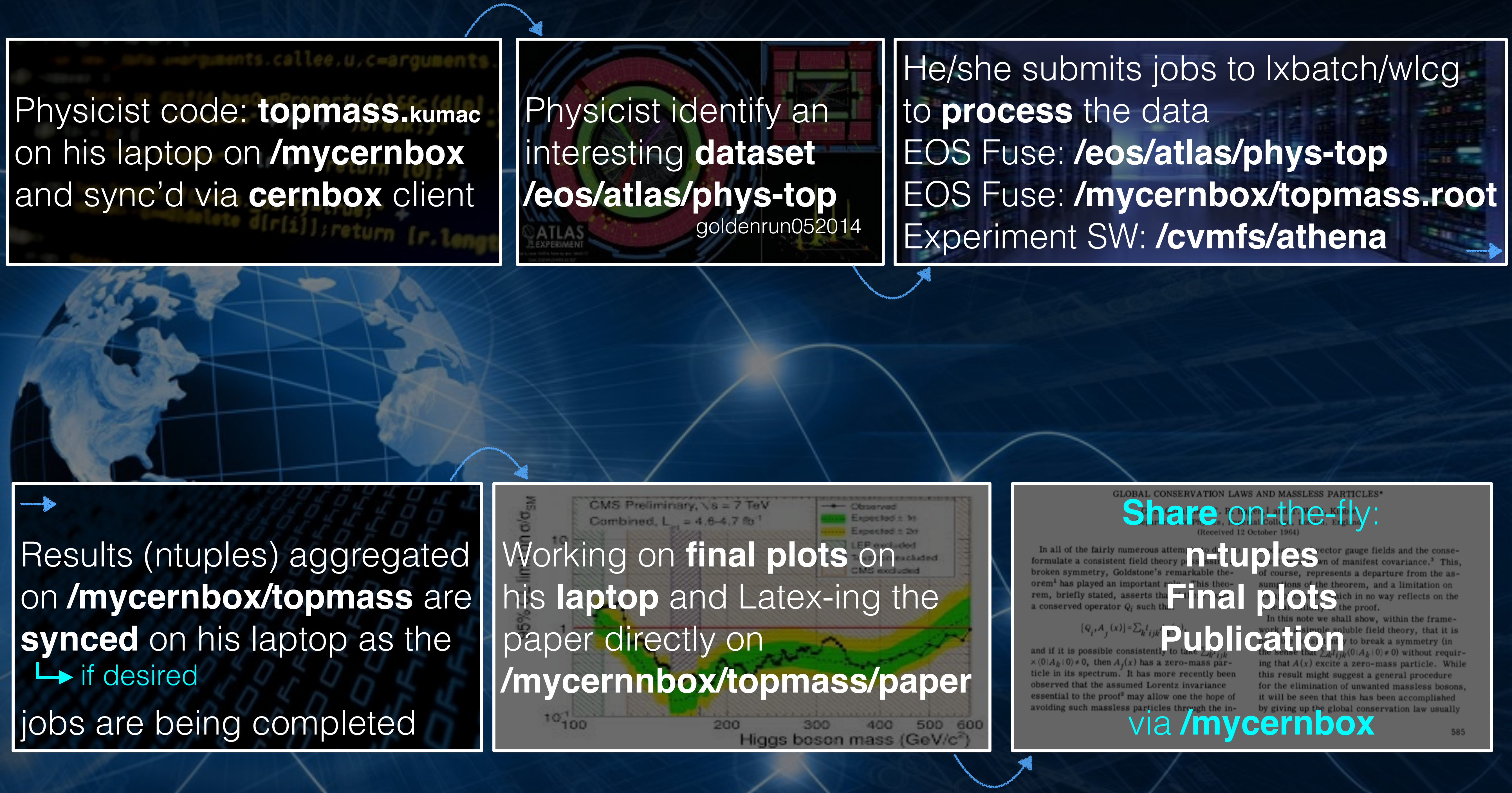
/eos/atlas

Data Access

/eos/atlas/topphys, /mycernbox, /cvmfs/athena

Goals

2 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.root**
Experiment SW: **/cvmfs/athena**

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

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

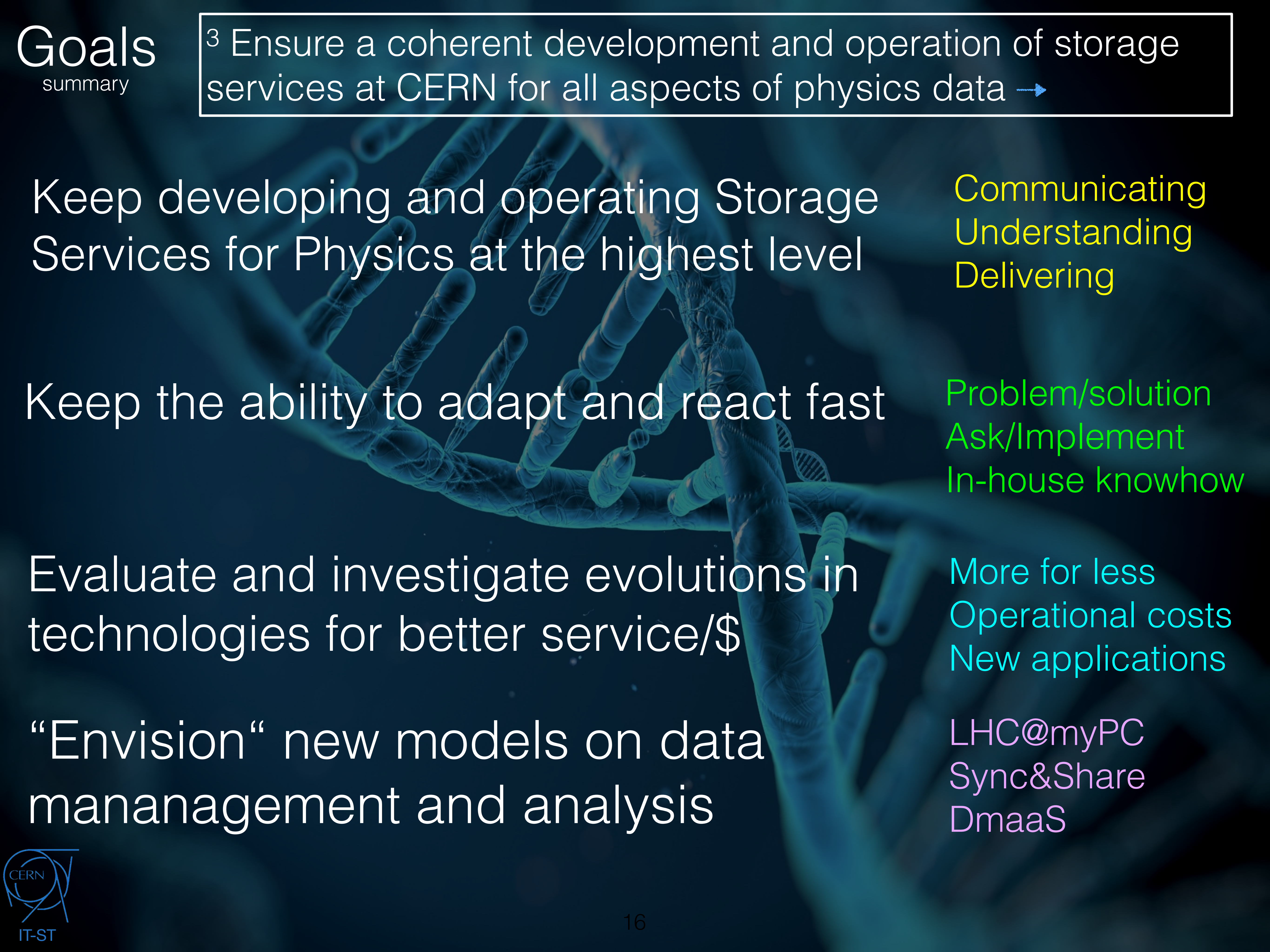
Share on-the-fly:
n-tuples
Final plots
Publication
via **/mycernbox**

is the enabling technology binding all this

Multi QoS Access patterns Protocols Redundancy

Goals
summary

[3] 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

3 We are here for you

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