CERN

CernVM-FS Operations in the CERN IT Storage Group

Dan van der Ster (CERN IT-ST)
CernVM Users Workshop 6-8 June 2016

Outline

- The new ops team
- Status in numbers
- Our (mostly virtual) architecture
- Evolving requirements
- Potential future architectures
- Client matters

Introductions

- CERN Stratum 0/1 and Squid Services now part of the IT Storage Group
  - CASTOR, EOS, AFS, NFS, Ceph, now CVMFS
- Team:
  - Dan van der Ster daniel.vanderster@cern.ch
  - Hervé Rousseau herve.rousseau@cern.ch
  - cvmfs-admins@cern.ch
- We inherited a flexible, clean service from Steve Traylen.

Stratum Zero Numbers

30 repositories across 19 release manager machines

Stratum Zeroes

Numbers according to ZFS

REPO	SIZE
aleph.cern.ch	594M
alice.cern.ch	**370G**
alice-ocdb.cern.ch	**1.1T**
ams.cern.ch	**2.5T**
atlasbuilds.cern.ch	709M
atlas.cern.ch	**1.1T**
bbp.epfl.ch	637M
belle.cern.ch	76G

REPO	SIZE
boss.cern.ch	25G
clicdp.cern.ch	384K
cms.cern.ch	**2.4T**
cms-opendata-conddb.cern.ch	60G
compass.cern.ch	512K
cvmfs-config.cern.ch	384K
delphi.cern.ch	12G
fcc.cern.ch	585M

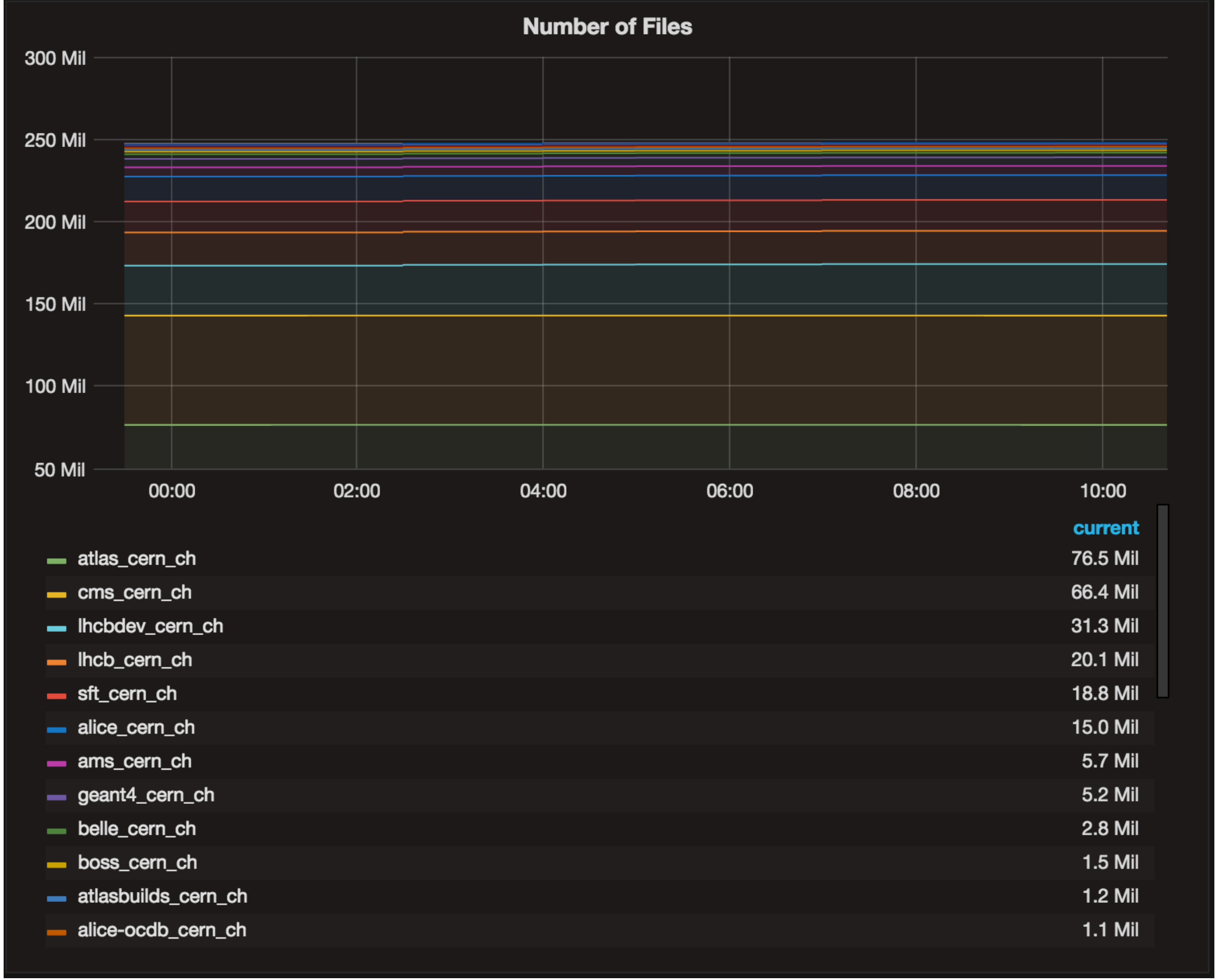
Stratum Zeroes

Numbers according to ZFS

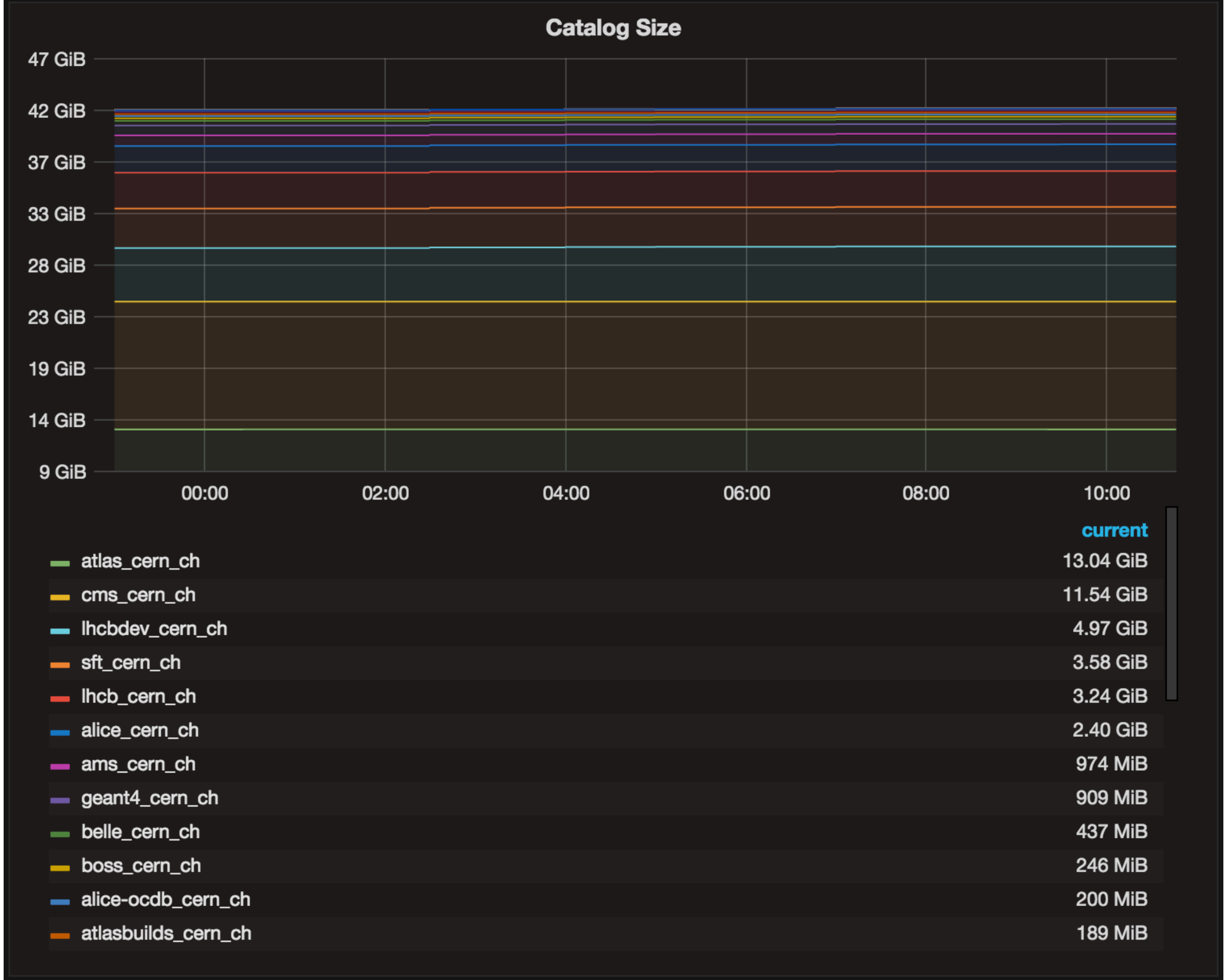
REPO	SIZE
ganga.cern.ch	1.1G
geant4.cern.ch	86G
grid.cern.ch	26G
lhcb.cern.ch	**1.1T**
lhcbdev.cern.ch	**742G**
moedal.cern.ch	512K
na49.cern.ch	392M

REPO	SIZE
na61.cern.ch	9.1G
na62.cern.ch	362M
opal.cern.ch	384K
sft.cern.ch	**492G**
test.cern.ch	31G
wlcg-clouds.cern.ch	384K

Plus a few zeroes not operated by CERN IT: cms-ib.cern.ch, atlas-nightlies.cern.ch, cernvm-prod.cern.ch



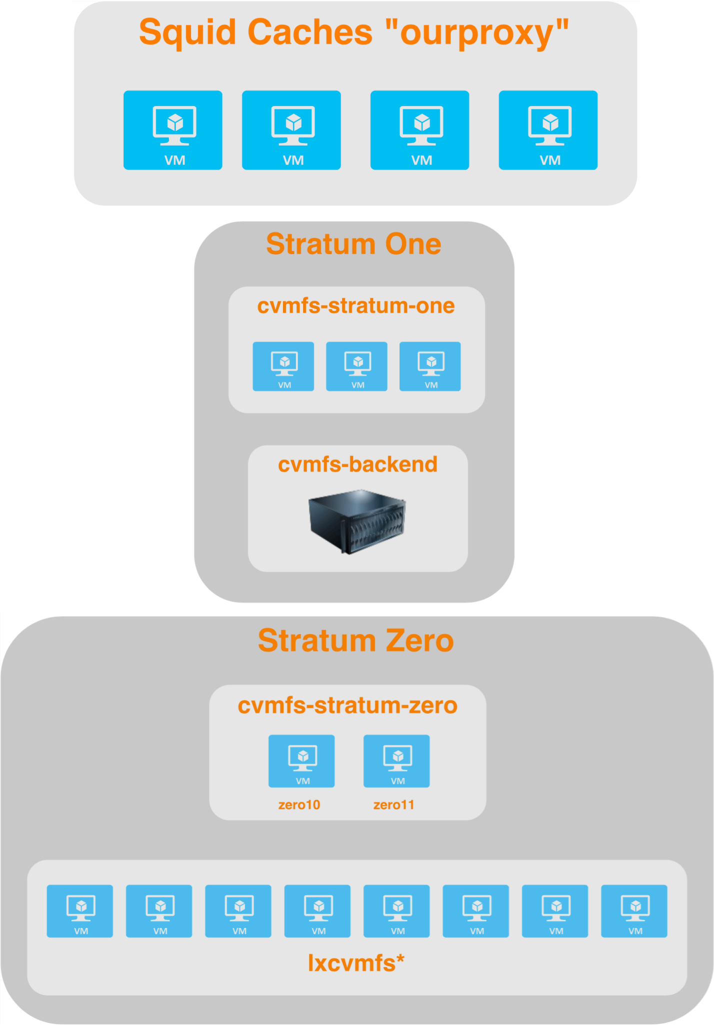
	current
atlas_cern_ch	76.5 Mil
cms_cern_ch	66.4 Mil
lhcbdev_cern_ch	31.3 Mil
lhcb_cern_ch	20.1 Mil
sft_cern_ch	18.8 Mil
alice_cern_ch	15.0 Mil
ams_cern_ch	5.7 Mil
geant4_cern_ch	5.2 Mil
belle_cern_ch	2.8 Mil
boss_cern_ch	1.5 Mil
atlasbuilds_cern_ch	1.2 Mil
alice-ocdb_cern_ch	1.1 Mil



Catalog Size

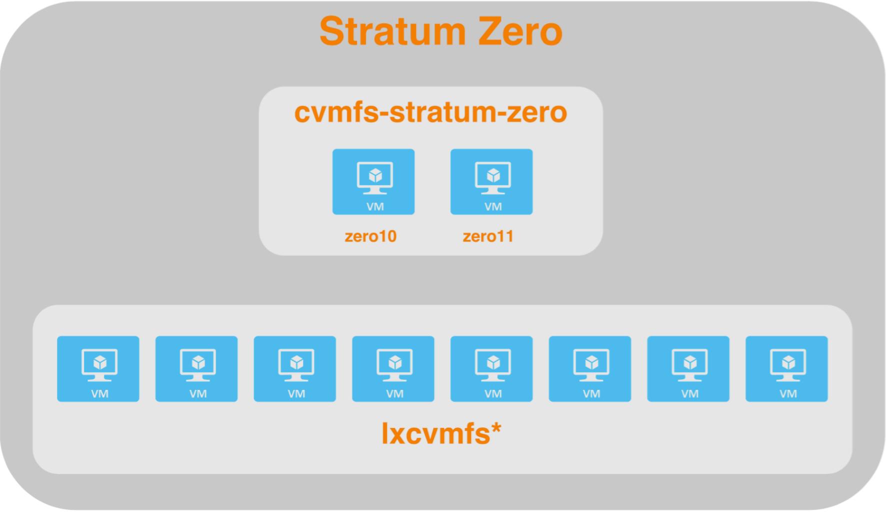
	current
atlas_cern_ch	13.04 GiB
cms_cern_ch	11.54 GiB
lhcbdev_cern_ch	4.97 GiB
sft_cern_ch	3.58 GiB
lhcb_cern_ch	3.24 GiB
alice_cern_ch	2.40 GiB
ams_cern_ch	974 MiB
geant4_cern_ch	909 MiB
belle_cern_ch	437 MiB
boss_cern_ch	246 MiB
alice-ocdb_cern_ch	200 MiB
atlasbuilds_cern_ch	189 MiB

Architecture



CERN Stratum Zero Architecture

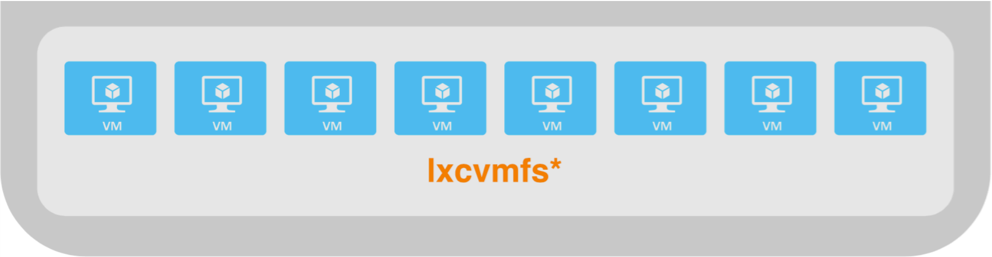
- Fully virtual stratum-0
- cvmfs-stratum-zero :
  - alias to best of a few small stateless VMs
- zero*.cern.ch:
  - Apache reverse proxy mod_proxy to hide the release manager machines.
- lxcvmfs*.cern.ch:
  - Aka cvmfs-<repo>.cern.ch
  - Large-ish VM with attached Ceph block storage
  - Release managers work here



Virtual Release Manager Machines

- /var/spool/cvmfs/<repo>.cern.ch is Ceph block storage
  - Flexible, reliable, durable
  - Tunable QoS a.k.a. IOPS/throughput
  - Thinly provisioned and resizable

- /var/spool/cvmfs/<repo>.cern.ch is ZFS
  - Snapshotting, incremental backups via replication to Wigner
  - Good performance, data integrity



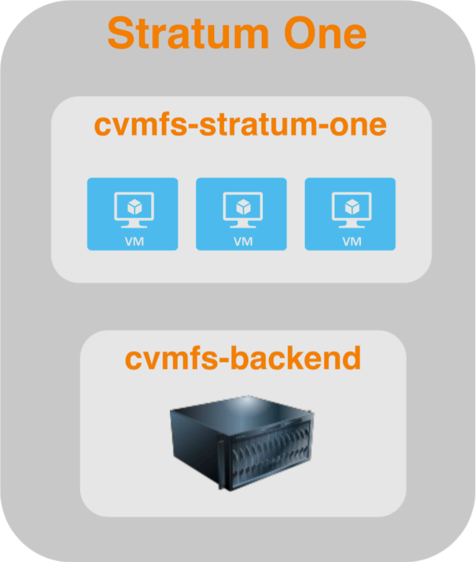
CERN Stratum One

- **cvmfs-backend.cern.ch**
  - Single large physical disk server: reliable and fast enough
  - Single large md raid10 ext4 fs, will need to expand it rather shortly

```
/dev/md5            11T  9.5T  743G  93% /srv
```

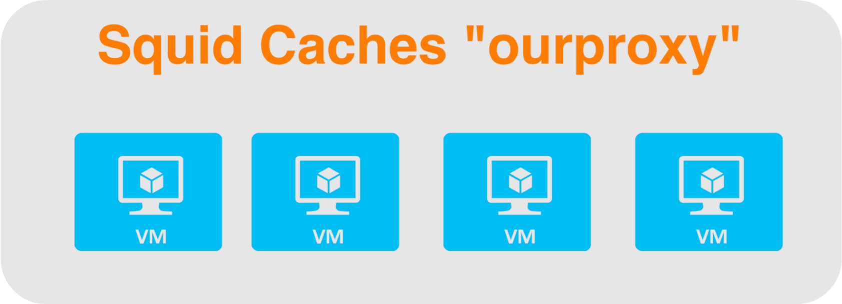
  - Is a *single point of failure*

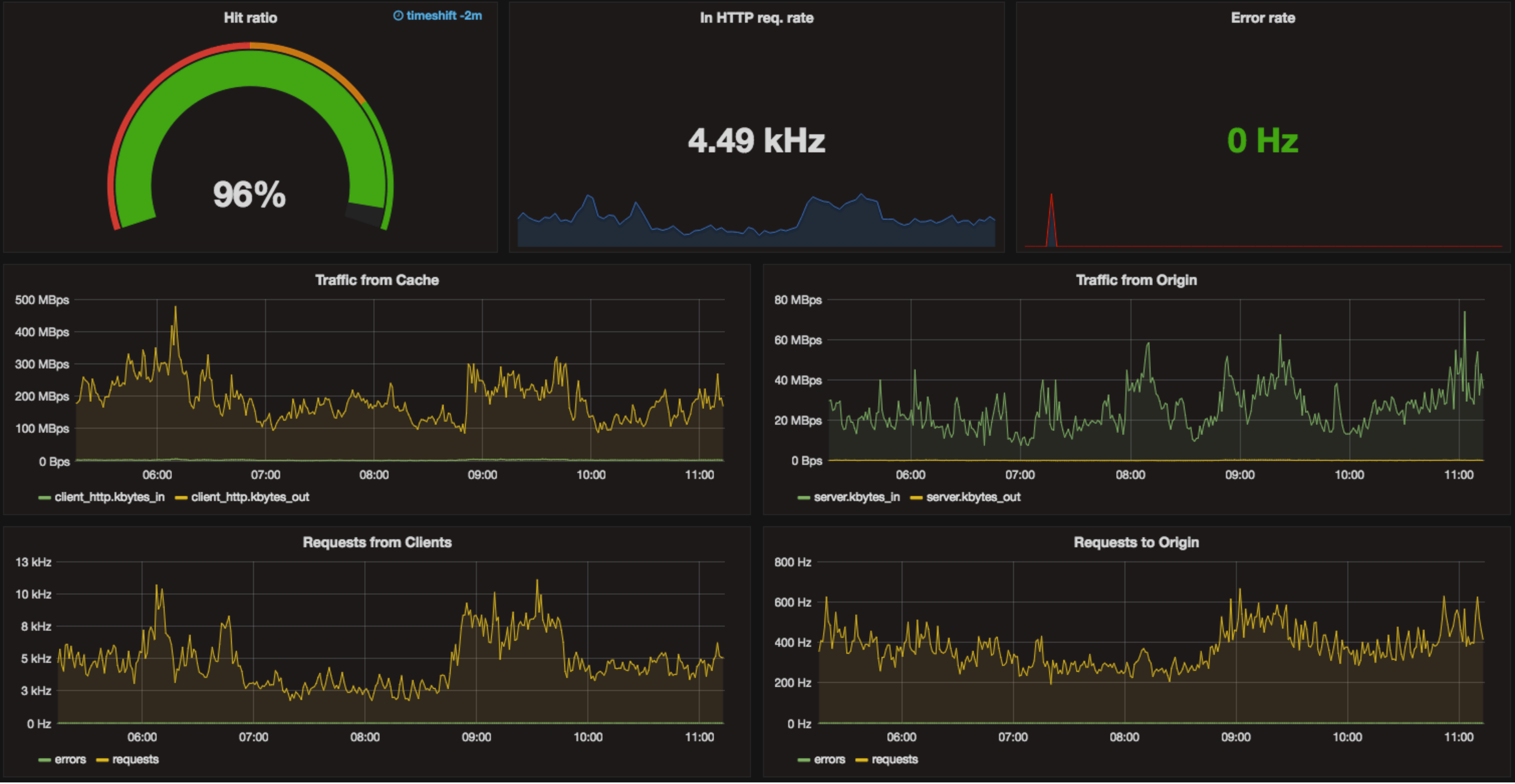
- **cvmfs-stratum-one.cern.ch**
  - Squids with around 300GB of attached Ceph volumes
  - (could grow the size of these)



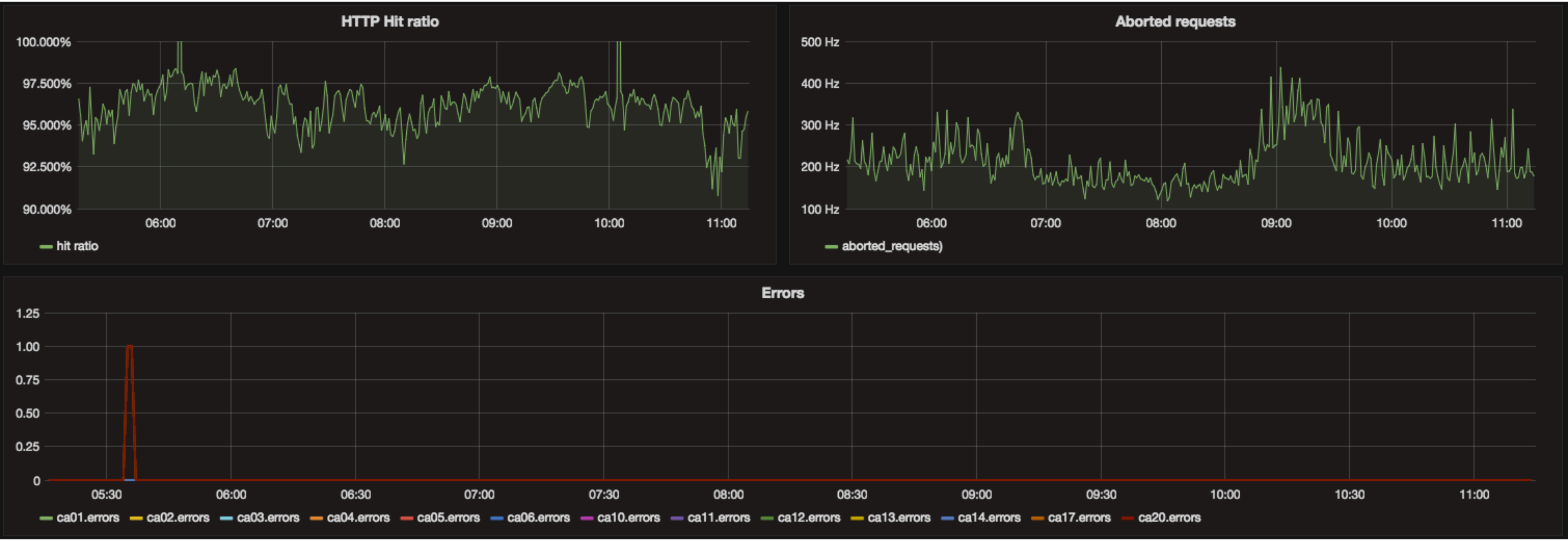
“OurProxy” Service

- Squids in Meyrin and Wigner with attached Ceph volumes (~200GB each)
- Quite reliable, no problems
- Working with Analysis WG to send all squid logs to HDFS
- In the meantime Hervé wrote a simple graphite probe to plot the squid_status metrics





https://filer-carbon.cern.ch/grafana/dashboard/db/squid-detailed



https://filer-carbon.cern.ch/grafana/dashboard/db/squid-detailed

Future

Moving out of AFS

- Motivated by the general decline in the community/project
- No hard deadline, but pushing things as much as we can.
  - Hoping to clean up during LS2 (2019)
- No single replacement product, but CVMFS is part of the solution.

- We expect growing CVMFS usage in the size and number of repositories
  - ATLAS already requested a ~20TB stratum 0 for nightlies
  - Total AFS project space for ATLAS is ~60TB (CMS around ~10TB)
  - Dedup in CVMFS will hopefully decrease this space requirement substantially

- There are >250 “project spaces” in AFS
  - Unclear which of these will end up in /cvmfs or /eos (or something else…)
  - Some repos will need restricted access – secure CVMFS development?

Pain points as we grow the service

- Too many repos:
  - Operating many more lxcvmfs* nodes will become a burden → need to scale up the machines to put more repos per node, or do something completely different (S3, NFS, CephFS, etc…)
  - Signing the whitelists is time consuming and error prone
    - Typing the same PIN 30 times twice a month ☹
  - The release manager nodes are almost equivalent to lxplus
    - Would be nice to separate the stratum zero servers from the interactive nodes
- Size of the repos:
  - Scaling the storage itself won’t be a problem at CERN
    - But puts new requirements on the stratum-1s
  - The “nightly build” use-cases will have a high rate of change
    - Maybe we shouldn’t replicate these to stratum-1s

Faster *Nightly* Stratum-Zeros

- Prototyping a new architecture for the nightlies use-cases

- Biggest VM we can get: 16 CPUs, 32GB RAM, local fast SSD
- Attach a huge Ceph volume (20TB in case of ATLAS)
- Use part of the local SSD as a ZFS write-ahead ZIL
- Disable *gc* at publish time (run every few days instead)
- Clients mount stratum-zero directly (perhaps via a squid)

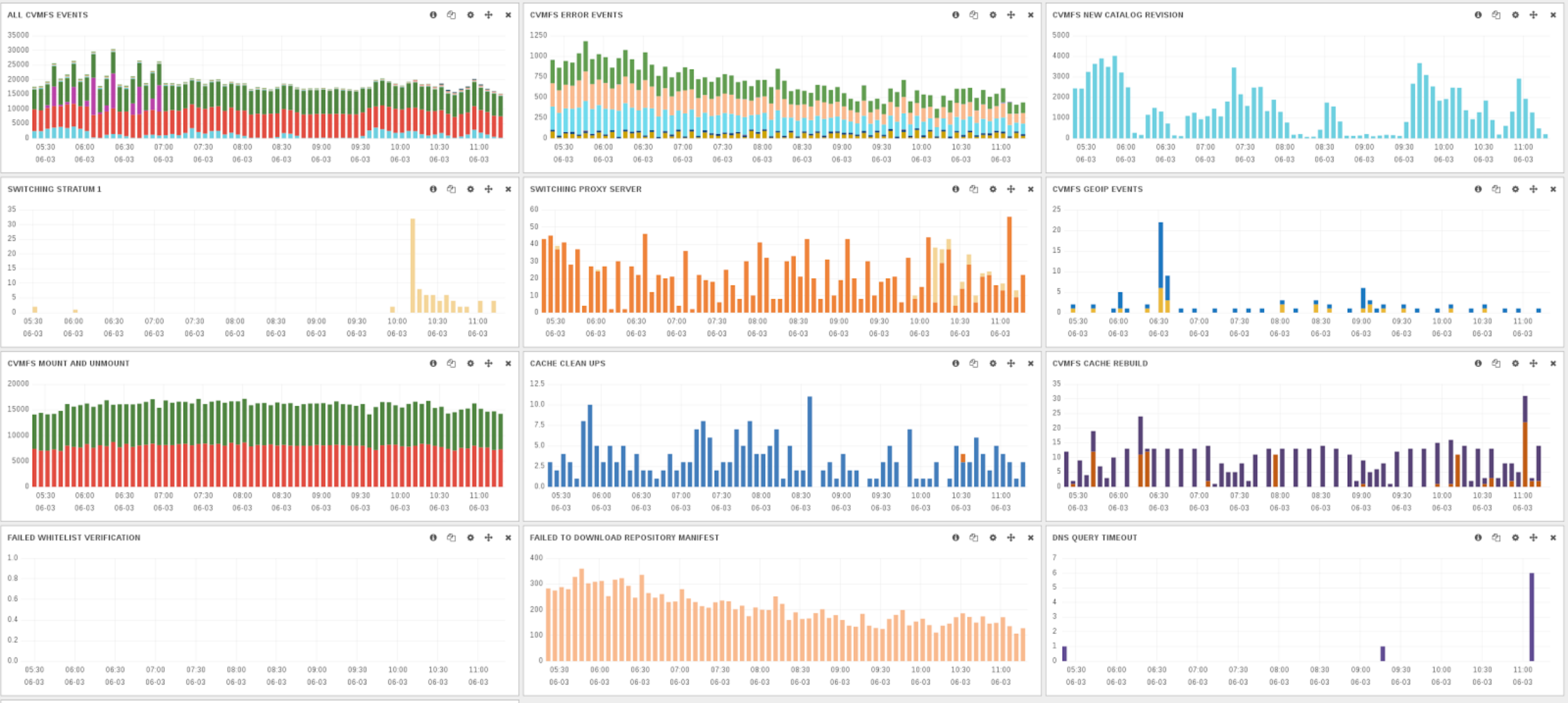
- If this doesn’t work, we’ll have to evaluate other backend storage tech (S3, CephFS, …)

Clients

CERNOPs-CvmFS Puppet Module

- Currently version 2.0.0, next version will deprecate all explicit hiera calls.
- Supports clients and stratum 0 well, stratum 1 support should be improved.
- Now used as part of puppet on desktop project at CERN
  - Desktops to get cvmfs easier
- https://forge.puppet.com/CERNOps/cvmfs

steve.traylen@cern.ch



Client Monitoring at CERN - syslog to ES & Kibana

steve.traylen@cern.ch

CVMFS on CERN Desktops

- A replacement of lcm (Local Configuration Manager) Quattor-based tool for CERN CentOS 7 is entering the early test stage.
- The new tool called - locmap - LO{cal) C(onfiguration) with MA(sterless) P(uppet) can be installed (and used) in parallel with existing lcm installation.
- It configures the same components as lcm using puppet modules, but we added a new cvmfs component.
- Ongoing:
  - Distribute CVMFS and locmap in standard CERN repositories. ETA: Q3
  - Replace lcm/ncm components by locmap/puppet modules for default CC7 installation.
  - ETA: when it's ready but not before CC7.3

```
# locmap --list

[Available Modules]

Module name : sudo [enabled]
Module name : sendmail [enabled]
Module name : ntp [enabled]
Module name : kerberos [enabled]
Module name : cvmfs [disabled]
Module name : ssh [enabled]
Module name : nscd [enabled]
Module name : afs [enabled]

# locmap --enable cvmfs
# locmap --configure cvmfs
```

thomas.oulevey@cern.ch

CVMFS Docker Volume Plugin

- Docker volume plugins are supported since 1.8
  - Providing integration with external storage systems
- CERN Cloud team provides a CVMFS plugin
  - https://gitlab.cern.ch/cloud-infrastructure/docker-volume-cvmfs

- Manages bind mounts between host and containers
  - Nicer interface on volume creation and deletion
- Packaging provided for CentOS7
- Instructions for Debian / Ubuntu

- Available by default in the CERN Cloud Container Service
  - http://clouddocs.web.cern.ch/clouddocs/containers/index.html

ricardo.rocha@cern.ch

CVMFS Docker Volume Plugin

- Creating a volume

```
> docker volume create -d cvmfs --name cms.cern.ch
cms.cern.ch
> docker volume ls
DRIVER             VOLUME NAME
cvmfs              cms.cern.ch
> docker volume rm cms.cern.ch
cms.cern.ch
```

- Launching a new container with a volume

```
> docker run -it --volume-driver cvmfs -v cms.cern.ch:/cms centos:7 /bin/bash
[root@874cbf8199d0 /]# ls /cms/
CMS@Home bootstrap_slc5_amd64_gcc462.log cmssw.git
...
```

ricardo.rocha@cern.ch

Conclusion

- New caretakers of CERN Stratum 0/1s + Squids
  - Ramping up our knowledge/experience of this service

- The (virtual) infrastructure runs well, and AFS decline means we’ll see growth in CVMFS
  - Investigating improvements in repo publishing speed as well as scalability (size + number of repos)

- Interesting work on client side to integrate with new platforms.

CERN