
CERN

CernVM
File system

Evolution of CernVM-FS Infrastructure at CERN

Enrico Bocchi
CERN IT – Storage

CernVM Users Workshop
3-5 June 2019, CERN, Geneva

CERN

Outline

- Introduction
- Stratum Zero
  - Numbers and size of the repositories
  - Virtualized infrastructure
  - S3 storage
  - CVMFS Gateway service
- Stratum One
- OurProxy caches
  - Dedicated Squids for selected repositories
- Future challenges

Introduction

- CVMFS Service inherited by IT Storage Group in 2016
  - Stratum Zero, Stratum One
  - Squid Caches (OurProxy service)
  - EOS, CERNBox, SWAN, CASTOR, AFS, NFS, Ceph, S3

- New team since Mid 2018:
  - Enrico Bocchi
  - Dan van der Ster
  - cvmfs-admins@cern.ch

Stratum Zero

Release Managers and Authoritative Storage

Stratum Zeros in Numbers

According to root catalog

- 49 repositories over 32 release managers
  - 43 repositories using Ceph block storage, 6 using S3 object storage
  - 25 release managers run SLC6, 7 run CC7
  - 2 additional CC7 machines for CVMFS Gateway Service
- 867.4 M files (less than 250 M in 2016)
  - 834.76 M on Ceph block storage
  - 32.63 M on S3 object storage
- 51.79 TB (~10.7 TB in 2016)
  - 49.94 TB on Ceph block storage (167.5 TB allocated, including backup)
  - 1.85 TB on S3 object storage (6 TB allocated)

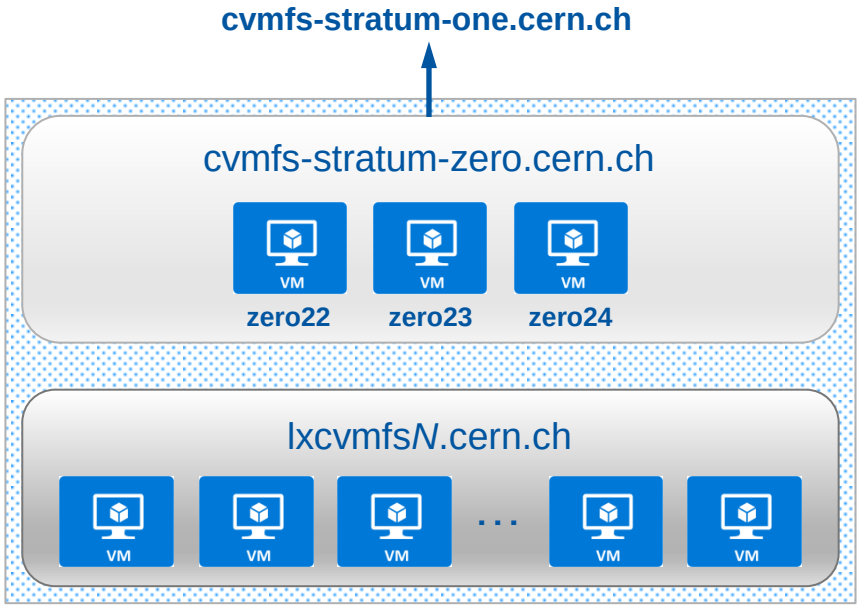
Stratum Zeros in Numbers

According to root catalog

Repository	Size [TB]	#Files [M]		Repository	Size [TB]	#Files [M]
atlas-nightlies.cern.ch	**8.53**	72.34		**lhcbdev-test.cern.ch** S3	1.46	25.96
cms.cern.ch	8.01	147.40		cms-ib.cern.ch	1.43	22.73
atlas.cern.ch	6.52	126.72		alice-nightlies.cern.ch	1.39	7.35
sft.cern.ch	6.26	**178.82**		belle.cern.ch	1.06	8.53
alice.cern.ch	4.69	31.79		sft-nightlies.cern.ch	0.74	102.90
ams.cern.ch	3.36	15.08		atlas-condb.cern.ch	0.72	0.01
lhcbdev.cern.ch	2.73	57.63		alice-ocdb.cern.ch	0.49	1.63
lhcb.cern.ch	2.40	37.69		clicdp.cern.ch	0.37	4.58

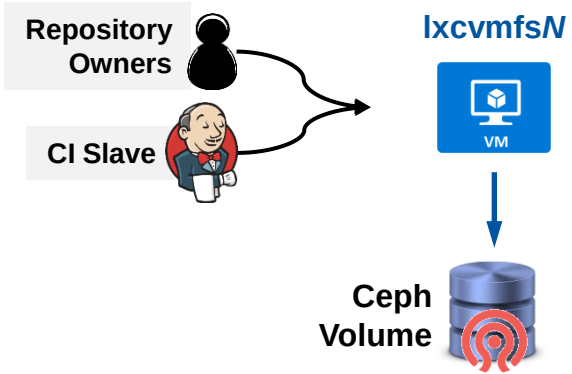
Stratum Zero Infrastructure

- Fully virtual on OpenStack
  - 304 CPUs, 547 GB memory
  - 32 release manager machines (lxcvmfs*N*.cern.ch)
  - 3 small reverse proxy machines (zero*N*.cern.ch)
- cvmfs-stratum-zero.cern.ch:
  - Alias to the best zero*N*.cern.ch reverse proxy
- zero*N*.cern.ch:
  - Reverse proxy (httpd) to hide the release manager machines
- lxcvmfs*N*.cern.ch:
  - `cvmfs_server …` lives here
  - Diverse flavors: small (1cpu,2GB) to 2xlarge (16,32)
  - Support for SLC6 and CC7
  - DNS aliased, e.g., cvmfs-config.cern.ch points to lxcvmfs54.cern.ch



Stratum Zero Storage

- 43 repositories use ZFS on Ceph volume
  - ZFS allows for snapshotting and incremental backups
  - ZFS quota and Ceph volume size adjustable on the needs
  - Ceph volume can be flexibly attached/detached to VMs

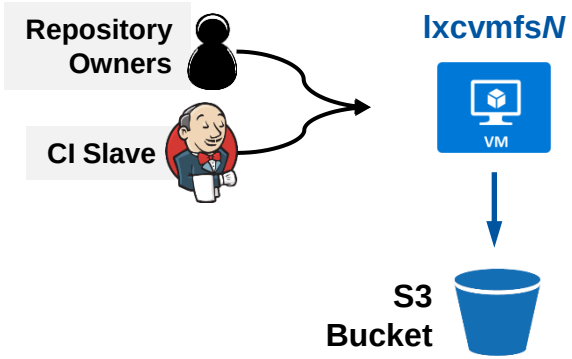


- Storage Layout

SLC6 AUFS	ZFS on Ceph volume mounted at /var/spool/cvmfs/<repo> ➢ /var/spool/cvmfs/<repo>/<repo>/data – Repo storage location ➢ /var/spool/cvmfs/<repo>/cv<repo> – $HOME for shared user
CC7 OvlFS	ZFS has two independent mount points ➢ /srv/cvmfs/<repo>/data – Repo storage location ➢ /home/cv<repo> – $HOME for shared user

Stratum Zero Storage

- 6 repositories use S3 storage
  - S3 is default production storage since Q4 2018
  - S3 quota handled on the S3 backend
  - One bucket per repo: cvmfs-<repo>
  - Storage endpoint s3.cern.ch with virtual hosting of buckets
  - Default params: 64 parallel connections, 10 retries



- Storage Layout
  - OpenStack Manila share (CephFS) for shared user home directory

CC7 OvlFS	➢ http://cvmfs-<repo>.s3.cern.ch/cvmfs/<repo> – Repo storage location ➢ /cephfs/cv<repo> – $HOME for shared user

S3 Storage

- Production service since 2018: s3.cern.ch
  - Originally used by ATLAS event service for ~3 years: up to 250TB used
- Single region radosgw cluster, ~2 PB
  - Load-balanced across 20 VMs with Traefik/RGW
  - 4+2 erasure coding for data, 3x replication for bucket indices
  - Now integrated with OpenStack Keystone for general service usage
- Future plans
  - Instantiation of a 2nd region: HW from Wigner + New HDDs
  - Demands for disk-only backup and disaster recovery are increasing
    E.g., CERNBox backup, Oracle databases backup

S3 Improvements

- Early 2019, upgrade cluster to BlueStore + bucket indices on SSD
  - Previous setup: 1x40GB SSD used as journal per 5-6 HDDs
  - Goal: Reuse the SSDs to keep BlueStore's RocksDB

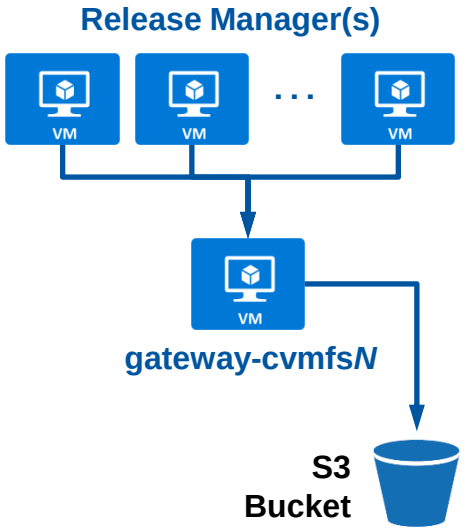
- Massive metadata performance increase:
  - Bucket indices in RocksDB on SSD
    is much faster than FileStore LevelDB on HDD
  - Metrics before were ~2kHz each!

- Sample workload: yum-reposync
  - From >2hr to ~1.2hr

Metric	Rate
PUT (new)	83kHz ± 4kHz
HEAD (not found)	63kHz ± 2kHz
DELETE	198kHz ± 15kHz

CVMFS Gateway

- Gateway service deployed on dedicated VMs for:
  - sw.hsf.org
  - atlas-nightlies.cern.ch (shadow of production atlas-nightlies)
- Sole access to authoritative storage
  - Release managers ship object packs to Gateway
  - Gateway commits changes to storage and updates repo manifest
- Running smoothly since Nov 2018 (on CC7 since Jan 2019)
  - Garbage collection: Only from Gateway, need root, stops Gateway service
  - Additional hop for change set from release manager to gateway machine
  - Not extensively tested with more repos // multiple release managers



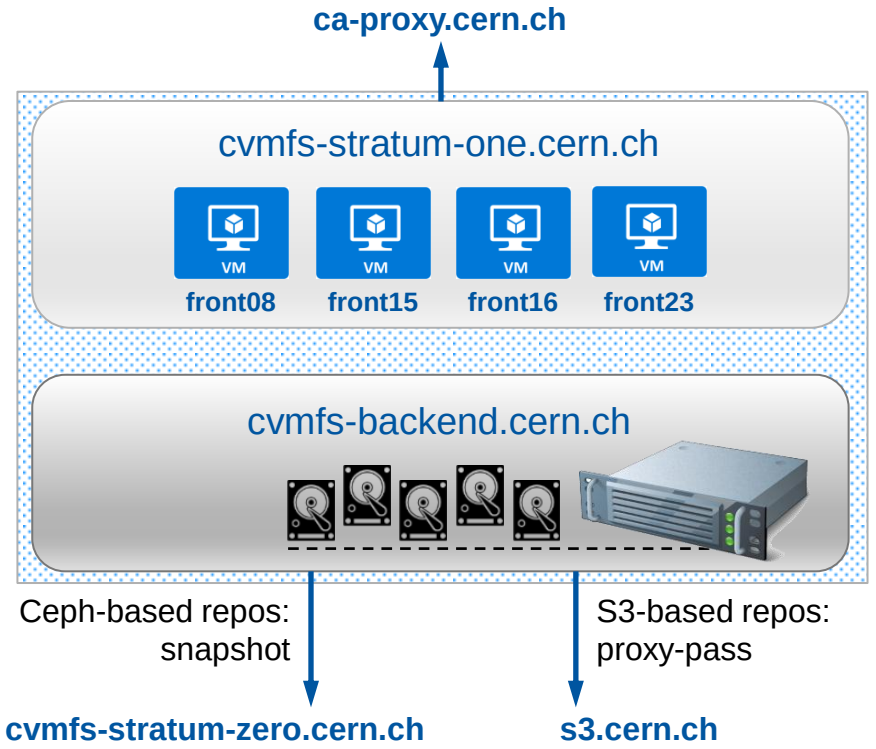
2

Stratum One

CERN Replica Servers and Front-line Caches

CERN Stratum One

- cvmfs-stratum-one.cern.ch:
  - Load-balanced alias to the front*N*.cern.ch squids
- front*N*.cern.ch:
  - Squid caches in front of cvmfs-backend.cern.ch
  - About 600GB of cache space (could grow)
  - Run httpd for statistics on clients
- cvmfs-backend.cern.ch:
  - Dedicated physical server with 24x6 TB disks
  - Single large ZFS volume of 130 TB
  - 250 GB L2ARC cache on SSD
  - Snapshots all Ceph-based repos
  - Proxy-passes to s3.cern.ch for S3-based repos

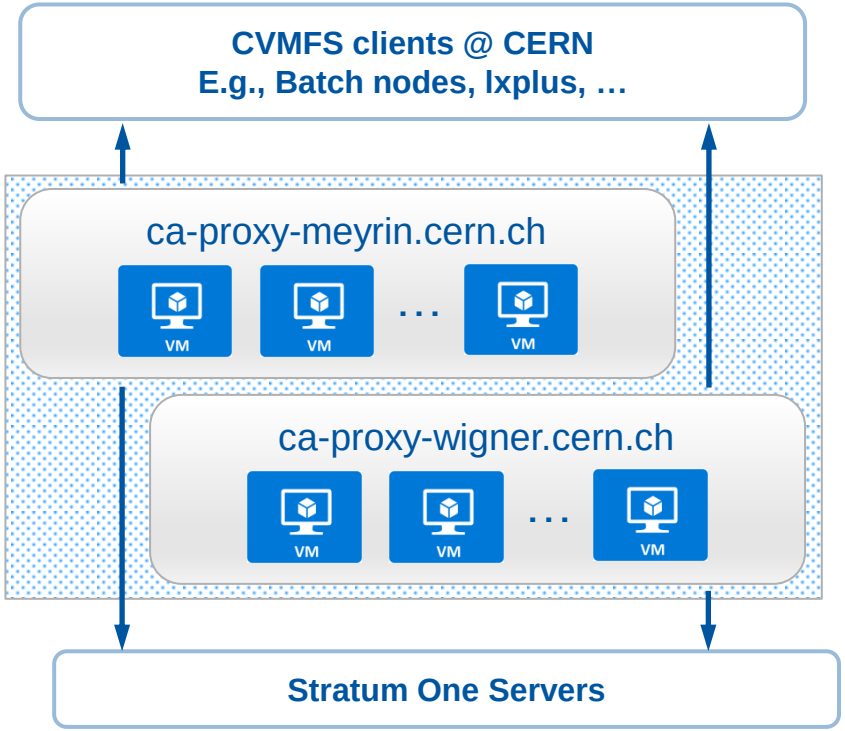


OurProxy

Squid Caches for Content Delivery

OurProxy

- ca-proxy.cern.ch:
  - Load-balanced alias for all caches at CERN
  - Union of ca-proxy-meyrin and ca-proxy-wigner
- ca-proxy-meyrin.cern.ch:
  - Load-balanced alias for caches in Meyrin
  - 14 VMs in total, 1000+ GB cache
- ca-proxy-wigner.cern.ch:
  - Load-balanced alias for caches in Wigner
  - 15 VMs in total, 800 GB cache
  - Decommissioned by End of 2019

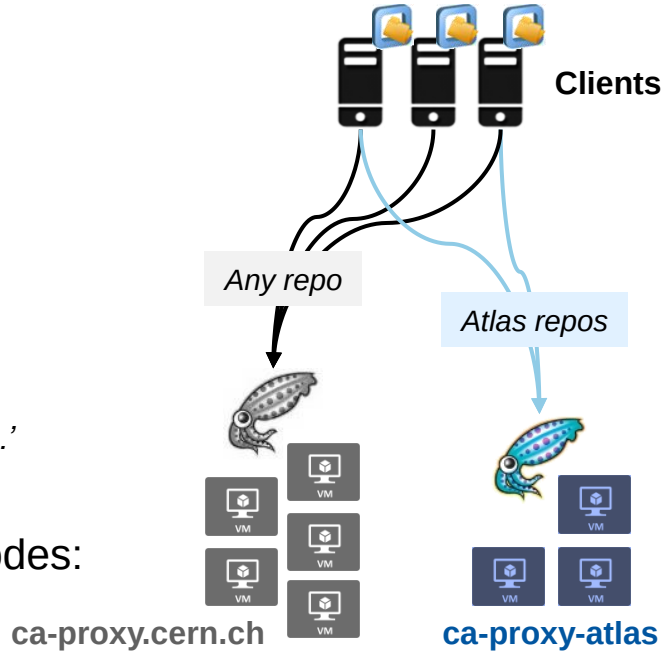


Dedicated Caches for ATLAS

- Deployment of dedicated squids
  - Reduce interference causing (potential) cache trashing
  - Improve cache utilization and hit ratio

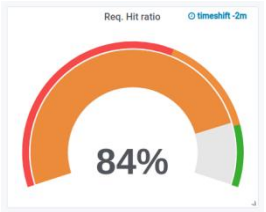
- ca-proxy-atlas.cern.ch:
  - Load-balanced alias to 4 VMs (600+ GB cache)
  - Serving traffic for {atlas, atlas-condb, atlas-nightlies, atlas-online-nightlies}.cern.ch
  - Client-side configuration – /etc/cvmfs/config.d/<repo>.local: *CVMFS_HTTP_PROXY='http://ca-proxy-atlas.cern.ch:3128, …'*

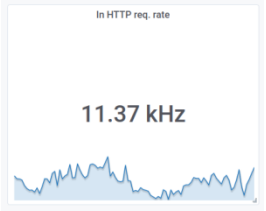
- Deployed on centrally managed batch and lxplus nodes:
  - May 15 to QA, May 22 merged to production
  - Early statistics look promising



OurProxy Monitoring

ca-proxy.cern.ch

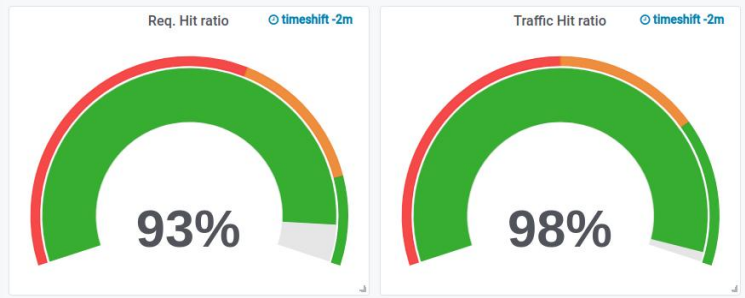


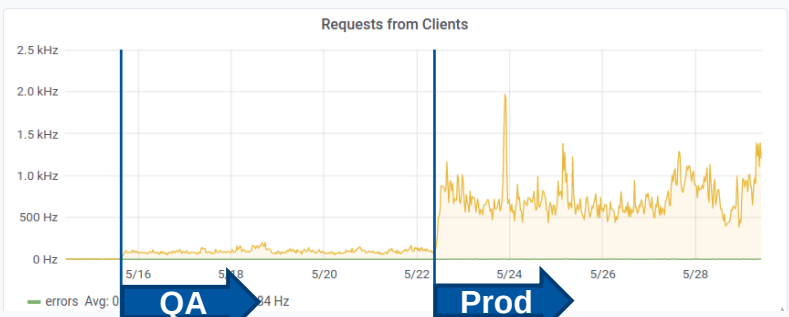


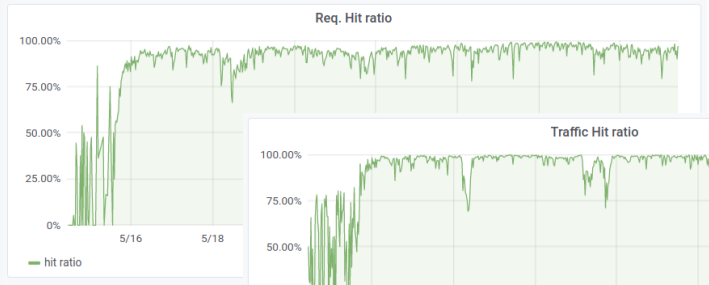
https://filer-carbon.cern.ch/grafana/d/000000063/squid-detailed

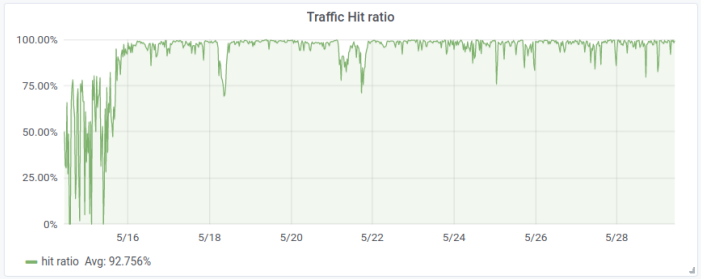
OurProxy Monitoring

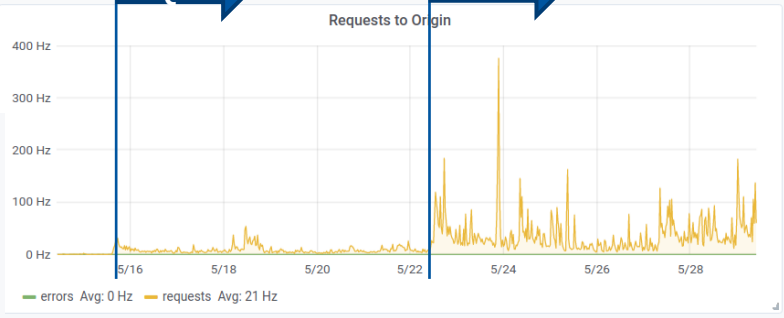
ca-proxy-atlas.cern.ch











https://filer-carbon.cern.ch/grafana/d/000000063/squid-detailed

4

Future Outlook

Challenges Ahead and Potential Solutions

Future Challenges (1)

- Many (repos and) release managers
  - Current model 1 repo ~= 1 release manager does not scale
  - Interventions (e.g., upgrade of cvmfs-server) are time consuming

  - Publishing through Gateway does not require persistent release managers
  - Release managers become stateless and disposable → Docker containers?

Future Challenges (2)

- Repository size growing
  - Provisioning of storage for Stratum Zeros is not a problem
  - Puts additional requirements on Squid caches and Stratum Ones
    e.g., risk of cache trashing, (much) bigger Stratum Ones

  - ✓ Dedicated sets of Squid reduce interference among repos
  - ✓ S3 storage infrastructure outperforms httpd on Stratum Zeros
    → Will CERN Stratum One still be needed?

Repository	2016 [TB]	2019 [TB]	Incr.
sft.cern.ch	0.49	6.26	1177 %
alice.cern.ch	0.37	4.69	1167 %
atlas.cern.ch	1.1	6.52	493 %
lhcbdev.cern.ch	0.74	2.73	269 %
cms.cern.ch	2.4	8.01	234 %

Future Challenges (3)

- Nightly-build use cases
  - Not completely solved yet
  - (Much) shorter publication and replication time

  - S3 storage and Gateway for parallel connections + concurrent publications
  - Do not replicate to Stratum One but serve directly from S3 (through a layer of Squids)

Conclusions

- Inherited stable service in Mid 2018
  - Still improving my knowledge and experience
- New features entered production
  - Support for SLC6 and **CC7**
  - Ceph volumes and **S3 storage**
  - **Gateway** service for S3-based repositories
- Future requirements
  - Reduce number of release manager machines for ease of management
  - Increase in storage volume coming from repository growth
  - Bigger and/or dedicated caches for content delivery


CERN