# Evolution of the S3 service at CERN

## as a storage backend for infrastructure services and software repositories

**Enrico Bocchi**

On behalf of the CEPH team

November 2019

CHEP, Adelaide

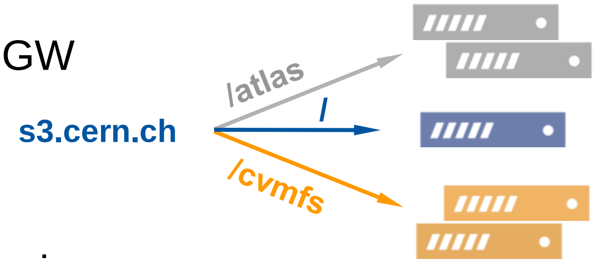
# S3 Service at CERN

- Recent achievements
- Future plans

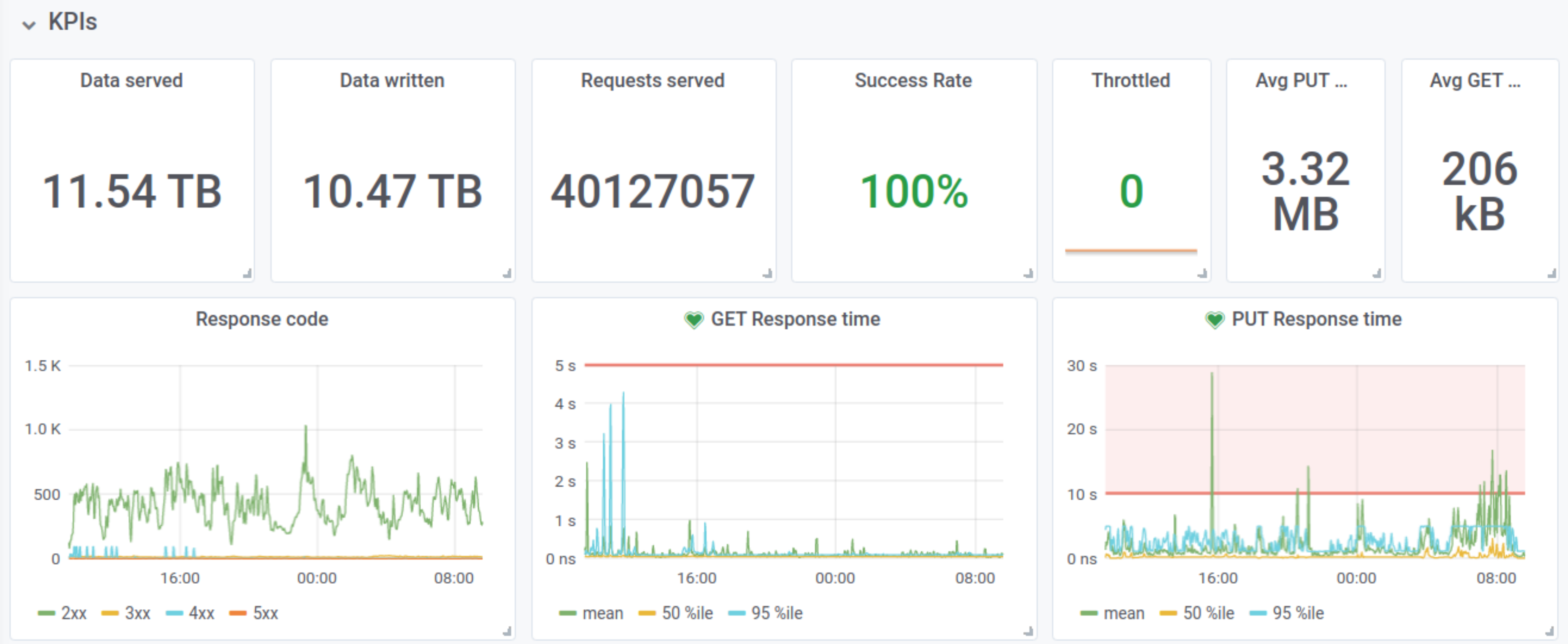
# S3 at CERN

- Production service since 2018: s3.cern.ch
  - ➢ Originally used by ATLAS event service for ~3 years: 275TB quota

- Single region RADOS Gateway cluster
  - ➢ 5000+ users, 2 PB raw capacity
  - ➢ 4+2 erasure coding for data, 3x replication for bucket indexes
  - ➢ Load-balanced across 16 VMs with Traefik / RadosGW
    - ✓ Dedicated RadosGW for specific use cases
    - ✓ 8x General Purpose, 4x CVMFS, 4x ATLAS
  - ➢ Integrated with OpenStack Keystone for general service usage



# One Day on RADOS Gateways



KPIs

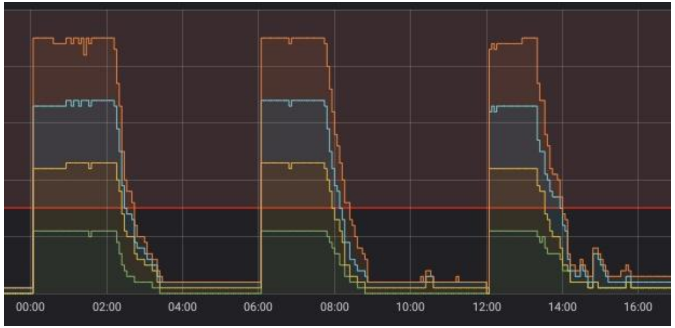
| Data served | Data written | Requests served | Success Rate | Throttled | Avg PUT ... | Avg GET ... |
|---|---|---|---|---|---|---|
| 11.54 TB | 10.47 TB | 40127057 | 100% | 0 | 3.32 MB | 206 kB |

# Achievements: BlueStore Upgrade

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

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

- Massive metadata performance increase
  - Bucket indexes 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



# Achievements: RadosGW Keystone Sync

- Integrate RadosGW authentication with OpenStack Keystone
  - OpenStack has a nice Object Store interface
  - Our users submit quota requests via the OpenStack Web UI

- Problem: Ceph-native integration with Keystone is slow
  - Each operation checks OpenStack Keystone for permission

- Solution: Synchronize Keystone credentials to RadosGW
  - OpenStack Mistral job writes the OpenStack credentials into RadosGW local users
  - Quota/Auth still managed by Keystone with local authentication performance

https://techblog.web.cern.ch/techblog/post/radosgw_sync_ec2_keys/

# Future plans

- Multi-region S3
  - Currently under evaluation
  - Second S3 region in CERN Prévessin (~5Km from main campus)
  - Objectives are high-availability and backup

2

# Applications of S3

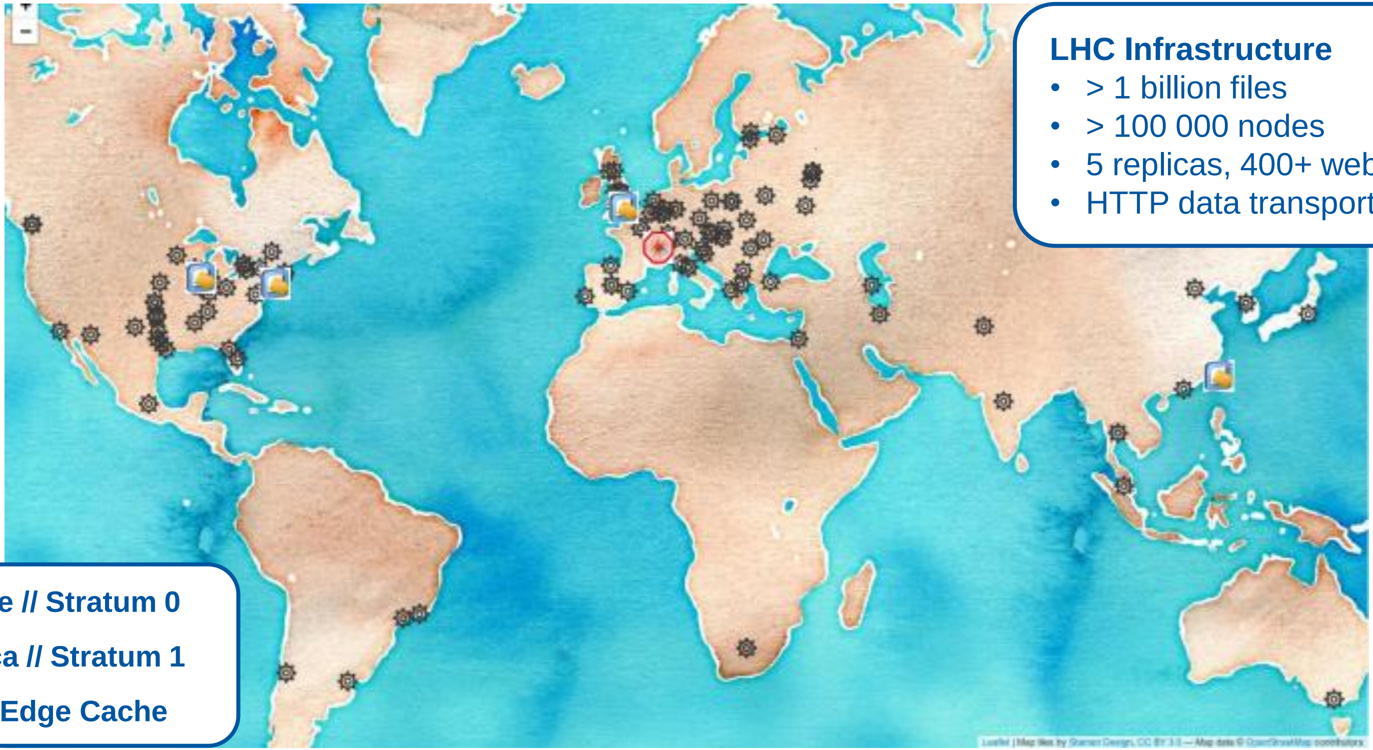
- Software distribution with CVMFS
- CERNBox backup to S3 via Restic

2

# Applications of S3

- **Software distribution with CVMFS**
- CERNBox backup to S3 via Restic

# The CernVM File System



Leaflet | Map tiles by Stamen Design, CC BY 3.0 — Map data © OpenStreetMap contributors

**LHC Infrastructure**

- > 1 billion files
- > 100 000 nodes
- 5 replicas, 400+ web caches
- HTTP data transport

**Source // Stratum 0**

**Replica // Stratum 1**

**Site // Edge Cache**

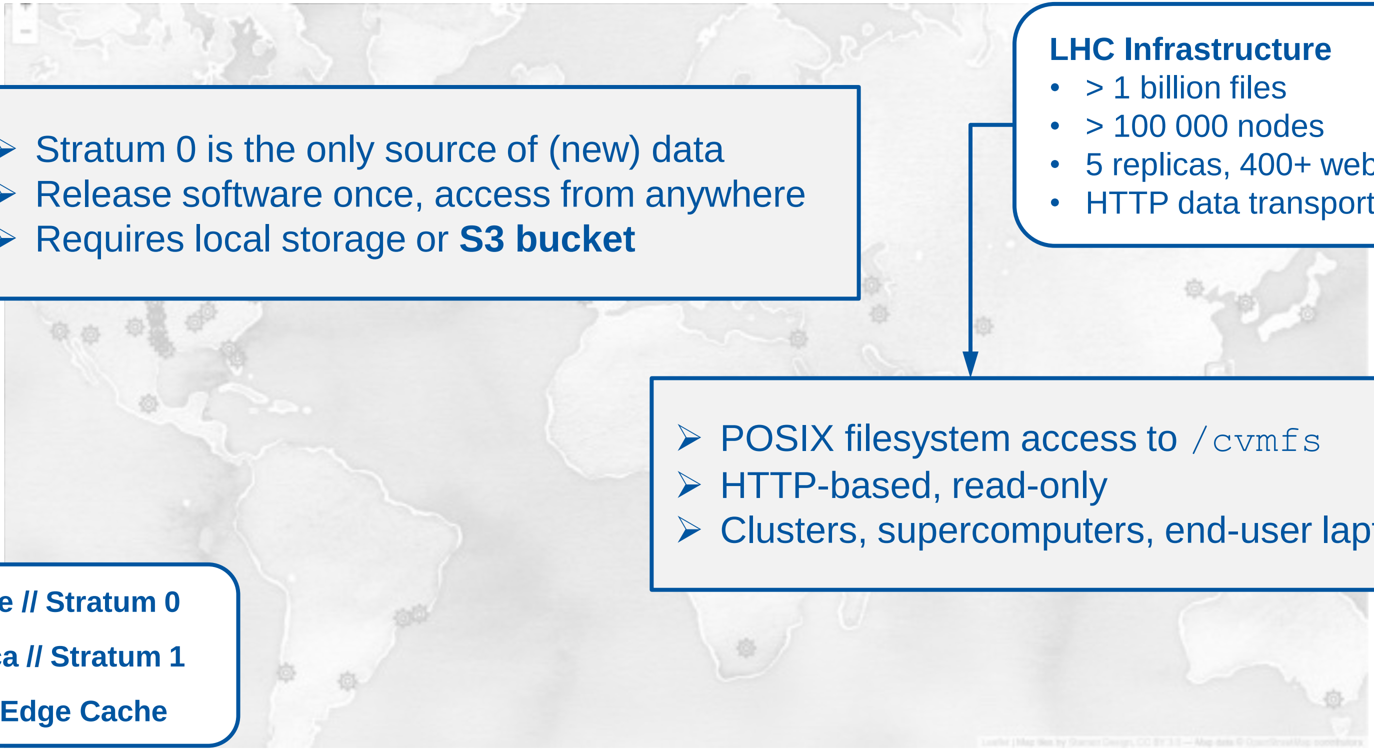
# The CernVM File System

- ➢ Stratum 0 is the only source of (new) data
- ➢ Release software once, access from anywhere
- ➢ Requires local storage or **S3 bucket**

**LHC Infrastructure**

- > 1 billion files
- > 100 000 nodes
- 5 replicas, 400+ web caches
- HTTP data transport

- ➢ POSIX filesystem access to `/cvmfs`
- ➢ HTTP-based, read-only
- ➢ Clusters, supercomputers, end-user laptop



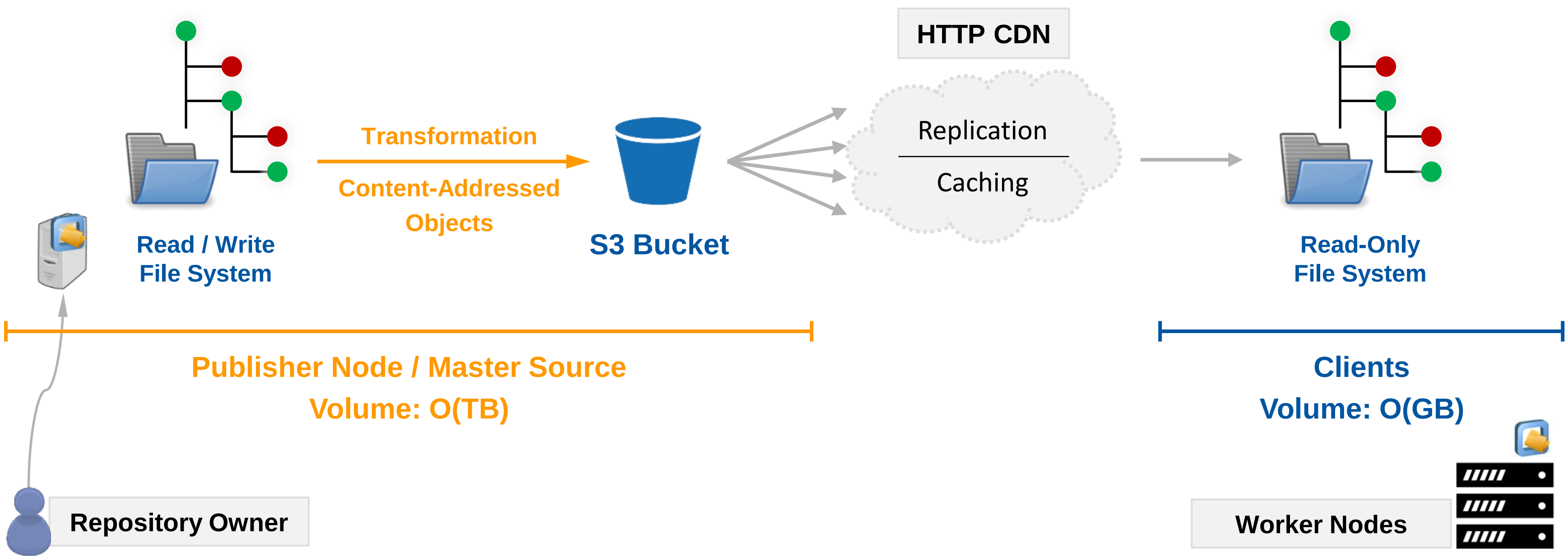
Leaflet | Map tiles by Stamen Design, CC BY 3.0 — Map data © OpenStreetMap contributors

**Source // Stratum 0**

**Replica // Stratum 1**

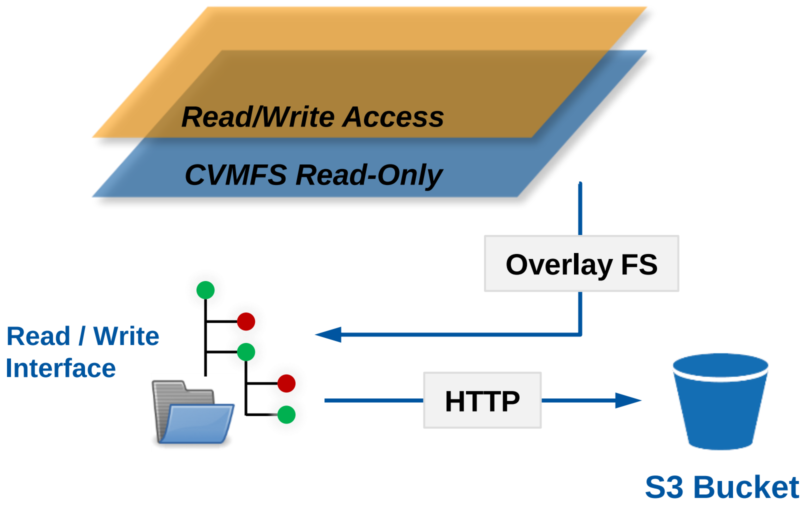
**Site // Edge Cache**

# S3 Object Store for CVMFS



# Publish on S3 with CVMFS

```
# cvmfs_server transaction myrepo.cern.ch
# cd /cvmfs/myrepo.cern.ch && tar xvf myarchive.tar.gz
# cvmfs_server publish myrepo.cern.ch
```



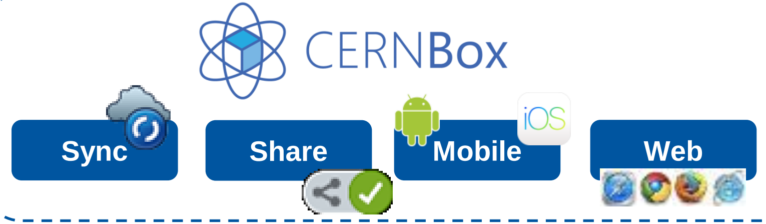
- Typical transaction workload
  - Bulk upload of O(100 k) files, 1 kB to 10 kB in size
  - ~5% of the files are new
  - (weekly) Garbage collection → Bulk delete of O(1M) files
  - Required throughput > 1 kHz using tens of HTTP streams

2

# Applications of S3

- Software distribution with CVMFS
- **CERNBox backup to S3 via Restic**

# CERNBox Backup Challenges



CERNBox

Sync

Share

Mobile

Web

- ➢ Available for all CERN user: 1 TB, 1 M files
- ➢ Ubiquitous file access: Web, mobile, sync to your laptop
- ➢ Not only physicists: engineers, administration, …

- Scalable backup solution
  - ➢ Stateless backup agents
  - ➢ Incremental backups, scattered in time
- Restore management and verification
  - ➢ On demand restore triggered by the user

| | 2017 | 2018 | 2019 | |
|---|---|---|---|---|
| **Users** | 8411 | 12686 | **18000** | *+41%* |
| **Files** | 176 Million | 470 Million | **1.1 Billion** | *+134%* |
| **Dirs** | 19 Million | 34 Million | **53 Million** | *+56%* |
| **Space Used** | 806 TB | 2.5 PB | **4 PB** | *+60%* |

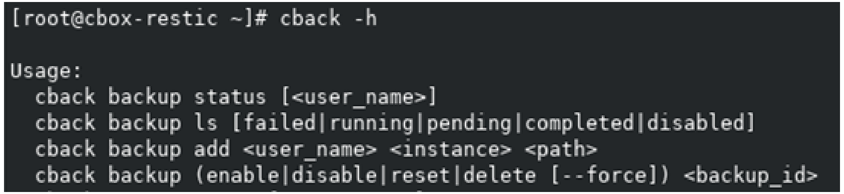
# Restic for CERNBox – cback

## Restic

- Efficient: File & chunk de-duplication, incremental backups
- Multiple backends (local, sftp, **S3**, Azure, Google Cloud, ... )

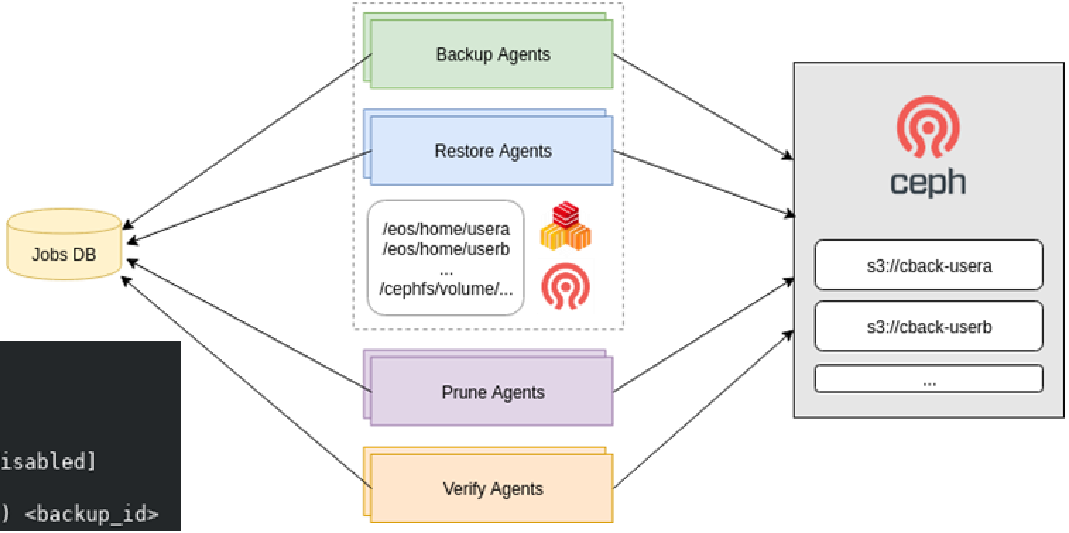
## cback

- Restic as a Service
- Backup jobs in MySQL DB
- CLI interface for management

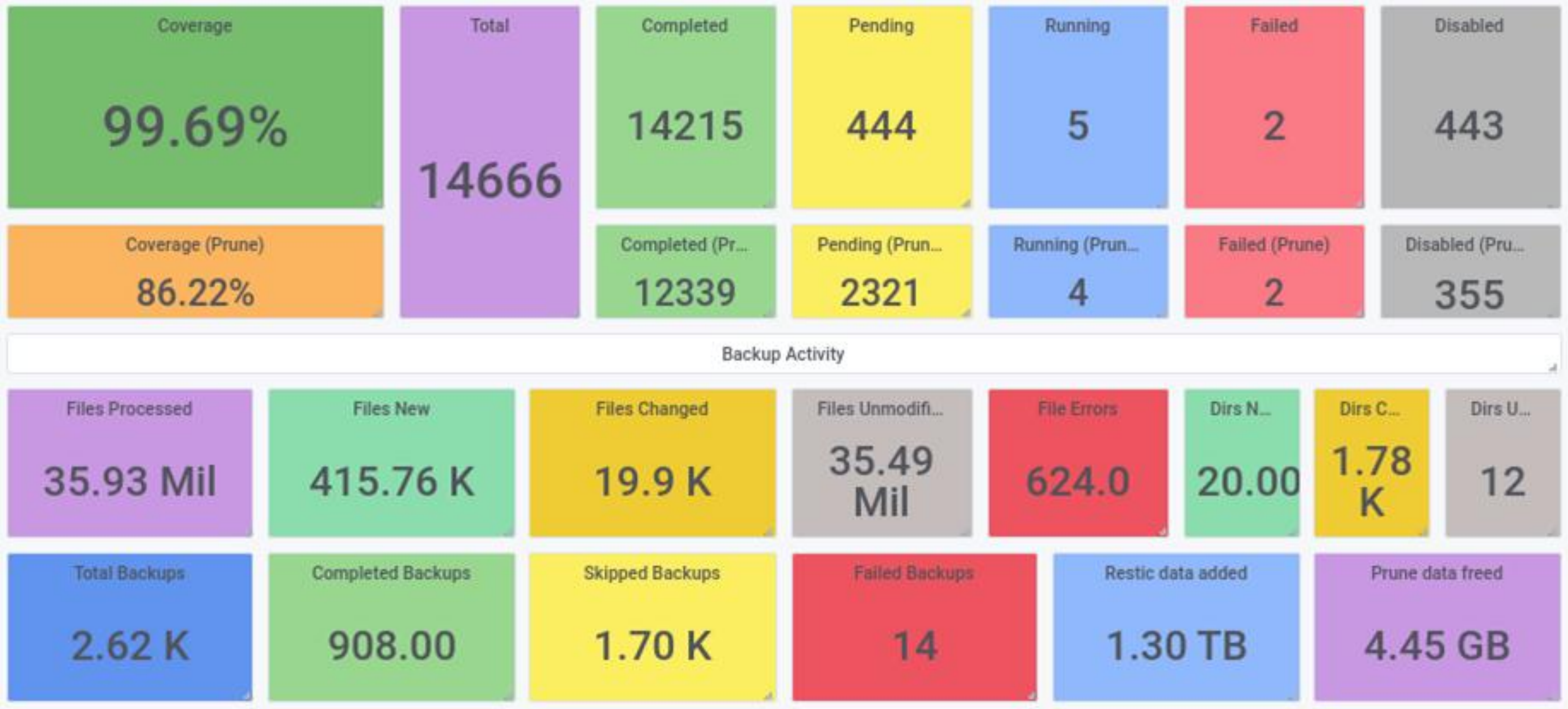


```
[root@cbox-restic ~]# cback -h

Usage:
  cback backup status [<user_name>]
  cback backup ls [failed|running|pending|completed|disabled]
  cback backup add <user_name> <instance> <path>
  cback backup (enable|disable|reset|delete [--force]) <backup_id>
```



# One Day on cback



Coverage
99.69%

Total
14666

Completed
14215

Pending
444

Running
5

Failed
2

Disabled
443

Coverage (Prune)
86.22%

Completed (Pr...
12339

Pending (Prun...
2321

Running (Prun...
4

Failed (Prune)
2

Disabled (Pru...
355

Backup Activity

Files Processed
35.93 Mil

Files New
415.76 K

Files Changed
19.9 K

Files Unmodifi...
35.49 Mil

File Errors
624.0

Dirs N...
20.00

Dirs C...
1.78 K

Dirs U...
12

Total Backups
2.62 K

Completed Backups
908.00

Skipped Backups
1.70 K

Failed Backups
14

Restic data added
1.30 TB

Prune data freed
4.45 GB

3

# Conclusions

# Conclusions

- S3 successful with diverse use cases
  - Stand-alone object storage (ATLAS event service, OpenStack end-users)
  - Storage backend for software distribution (CVMFS)
  - Backup and recovery solution for other storage services (CERNBox)

- Future improvements
  - Planning deployment of second S3 region
  - CVMFS would benefit from bundled-request capability
    e.g., multi-HEAD, multi-PUT to reduce latency

# Evolution of the S3 service at CERN

## as a storage backend for infrastructure services and software repositories

**Thank you!**

Enrico Bocchi

enrico.bocchi@cern.ch

CERN

# Outline

- S3 service at CERN
  - Recent achievements
  - Future plans

- S3 use cases
  - Distribution of HEP software with CVMFS
  - CERNBox Backup with Restic and `cback`

- Conclusions

# The CernVM File System

https://github.com/cvmfs/cvmfs

## Write

- A publish-subscribe file system tuned for maximum dissemination

```
$ cvmfs_server transaction myrepo.cern.ch
$ cvmfs_server publish myrepo.cern.ch
```

- Publisher node is the single source of (new) data: read-write permissions
- Install applications once on the publisher, access from anywhere

## Read

- POSIX file system access to globally available directory `/cvmfs`

```
 $ ls /cvmfs/myrepo.cern.ch
myFOLDER  myREADME.md
```

- HTTP-based read-only access
- RedHat, Debian, Ubuntu, macOS, …
- Clusters, cloud, supercomputers, end-user laptop