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

Response code and Response time graphs are also shown.
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

- 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

<table>
<thead>
<tr>
<th>Metric</th>
<th>Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>PUT (new)</td>
<td>83kHz ± 4kHz</td>
</tr>
<tr>
<td>HEAD (not found)</td>
<td>63kHz ± 2kHz</td>
</tr>
<tr>
<td>DELETE</td>
<td>198kHz ± 15kHz</td>
</tr>
</tbody>
</table>
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
Applications of S3

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

- Software distribution with CVMFS
- CERNBox backup to S3 via Restic
The CernVM File System

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

Source // Stratum 0
Replica // Stratum 1
Site // Edge Cache
S3 Object Store for CVMFS

Publisher Node / Master Source
Volume: O(TB)

Repository Owner

Transformation
Content-Addressed Objects

S3 Bucket

HTTP CDN
Replication
Caching

Read-Only File System

Clients
Volume: O(GB)

Worker Nodes

Read / Write File System

Volume: O(GB)
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
Applications of S3

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

- Scalable backup solution
  - Stateless backup agents
  - Incremental backups, scattered in time

- Restore management and verification
  - On demand restore triggered by the user

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

<table>
<thead>
<tr>
<th></th>
<th>2017</th>
<th>2018</th>
<th>2019</th>
<th>Change</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Users</strong></td>
<td>8411</td>
<td>12686</td>
<td>18000</td>
<td>+41%</td>
</tr>
<tr>
<td><strong>Files</strong></td>
<td>176 Million</td>
<td>470 Million</td>
<td>1.1 Billion</td>
<td>+134%</td>
</tr>
<tr>
<td><strong>Dirs</strong></td>
<td>19 Million</td>
<td>34 Million</td>
<td>53 Million</td>
<td>+56%</td>
</tr>
<tr>
<td><strong>Space Used</strong></td>
<td>806 TB</td>
<td>2.5 PB</td>
<td>4 PB</td>
<td>+60%</td>
</tr>
</tbody>
</table>
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

<table>
<thead>
<tr>
<th>Coverage</th>
<th>Total</th>
<th>Completed</th>
<th>Pending</th>
<th>Running</th>
<th>Failed</th>
<th>Disabled</th>
</tr>
</thead>
<tbody>
<tr>
<td>99.69%</td>
<td>14666</td>
<td>14215</td>
<td>444</td>
<td>5</td>
<td>2</td>
<td>443</td>
</tr>
<tr>
<td>Coverage (Prune)</td>
<td>86.22%</td>
<td>12339</td>
<td>2321</td>
<td>4</td>
<td>2</td>
<td>355</td>
</tr>
</tbody>
</table>

## Backup Activity

<table>
<thead>
<tr>
<th>Files Processed</th>
<th>Files New</th>
<th>Files Changed</th>
<th>Files Unmodified</th>
<th>File Errors</th>
<th>Dirs N</th>
<th>Dirs C</th>
<th>Dirs U</th>
</tr>
</thead>
<tbody>
<tr>
<td>35.93 Mil</td>
<td>415.76 K</td>
<td>19.9 K</td>
<td>35.49 Mil</td>
<td>624.0</td>
<td>20.00</td>
<td>1.78 K</td>
<td>12</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Total Backups</th>
<th>Completed Backups</th>
<th>Skipped Backups</th>
<th>Failed Backups</th>
<th>Restic data added</th>
<th>Prune data freed</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.62 K</td>
<td>908.00</td>
<td>1.70 K</td>
<td>14</td>
<td>1.30 TB</td>
<td>4.45 GB</td>
</tr>
</tbody>
</table>
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
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

  ```
  $ cmvms_server transaction myrepo.cern.ch
  $ cmvms_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

---

[S3 Evolution at CERN](#)