



Towards a Serverless CernVM-FS

Jakob Blomer, for the CernVM-FS team
CERN, EP-SFT
CHEP 2018, Sofia

CernVM-FS Scale of Deployment



- > 1 billion files under management
- Mission critical for the large HEP experiments
- > 100 repositories within and beyond physics
- WLCG: 5 mirror servers and 400 web caches (shared with Frontier)



- 1 Provide uniform, consistent, and versioned POSIX file system access to /cvmfs

```
λ ls /cvmfs/cms.cern.ch
slc7_amd64_gcc700  slc7_ppc64le_gcc530  slc7_aarch64_gcc700  slc6_mic_gcc481
...
```

on **grids**, **clouds**, **supercomputers** and **end user laptops**

read

publish

- 2 Populate and propagate new and updated content
 - Support an increasing number of /cvmfs/...writers



- 1 Provide uniform, consistent, and versioned POSIX file system access to /cvmfs

```
λ ls /cvmfs/cms.cern.ch  
slc7_amd64_gcc700  slc7_ppc64le_gcc530  slc7_aarch64_gcc700  slc6_mic_gcc481  
...
```

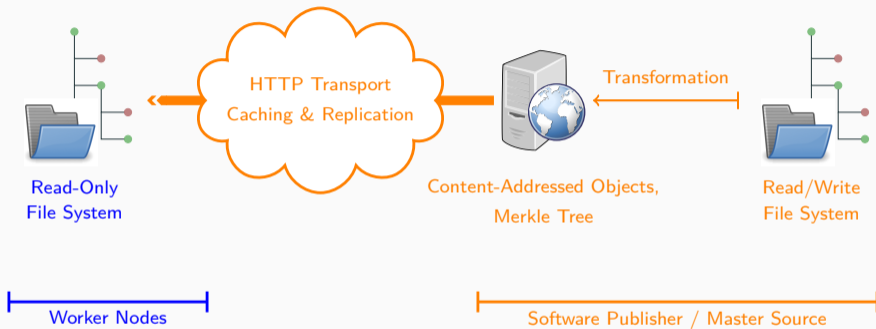
on **grids**, **clouds**, **supercomputers** and **end user laptops**

read

publish

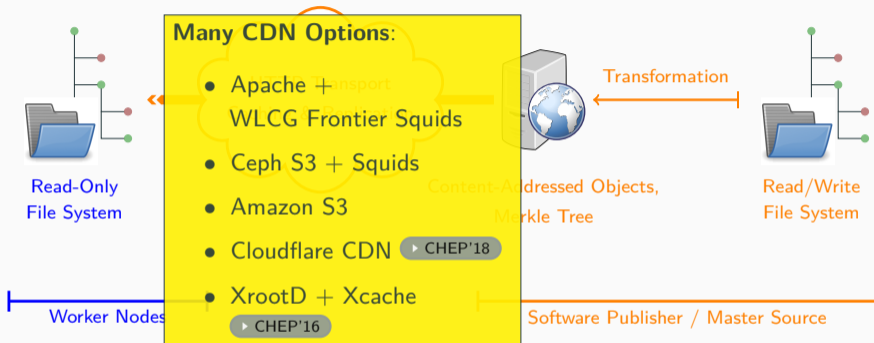
- 2 Populate and propagate new and updated content
 - Support an increasing number of /cvmfs/... writers

Propagation delay
down to 5–10 min



- Reading and writing treated asymmetrically
- Immutable objects, stateless services

- HTTP transport
- Aggressive caching



- Reading and writing treated asymmetrically
- Immutable objects, stateless services

- HTTP transport
- Aggressive caching



📄 13,453 commits

🌿 68 branches

📦 37 releases

👤 31 contributors

📄 BSD-3-Clause

- R&D character attracted excellent engineers; some went on to Mesosphere, Amazon, Pivotal, etc.
- Contributions by OpenLab, CERN KT fund, CERN IT, FNAL, U Nebraska, GSoC, industry
- User workshops at RAL and CERN
- **Active user, operations, and code community**

-
- 2008 ··· ● CernVM R&D: 1st prototype.
 - 2010 ··· ● CernVM-FS @ CHEP Taipei.
 - 2011 ··· ● Code moved to Github.
 - 2012 ··· ● Picked up by PIC, RAL; grid deployment task force.
 - 2014 ··· ● DPHEP / software preservation.
 - 2016 ··· ● Growing outside interest: LIGO, EUCLID, Mesosphere,
 - 2017 ··· ● Deployment on HPC.



① Production Software

Example: [/cvmfs/ligo.egi.eu](#)

- ✓ Most mature use case
- 🔥 Continuous effort on HPC systems

③ Extracted Container Images

Example: [/cvmfs/singularity.opensciencegrid.org](#)

- ✓ Works out of the box with Singularity
- ✓ CernVM-FS driver for Docker ▶ ACAT'17
- 🔥 Easy ingestion of images ▶ Prototype

② Integration Builds

Example: [/cvmfs/lhcbdev.cern.ch](#)

- ✓ High churn, requires regular garbage collection
- 🔥 Instant update propagation to file system clients ▶ CHEP'18

④ Auxiliary data

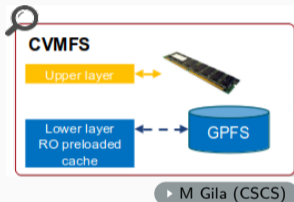
Example: [/cvmfs/alice-ocdb.cern.ch](#)

- ✓ Benefits from internal versioning
- Depending on volume requires more planning for the CDN components

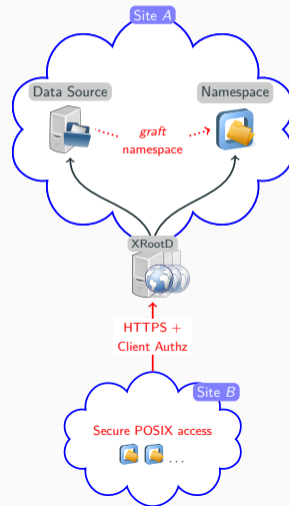


① HPC Client Deployment

- Piz Daint:
Europe's fastest
supercomputer
- Runs ATLAS,
CMS, and LHCb
jobs from native
Fuse client



② Data Namespace: `/cvmfs/*.osgstorage.org`

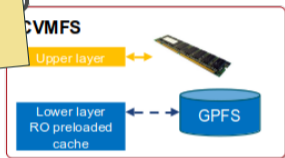




1 HPC Client Deployment

Multi-layer cache +
memory cache plugin

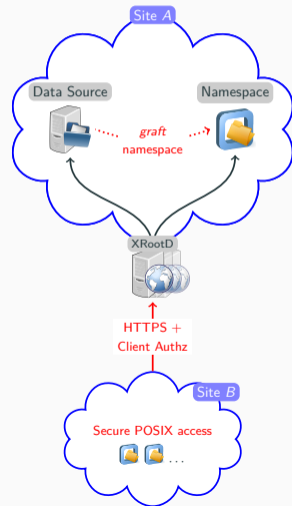
- Runs ATLAS, CMS, and LHCb jobs from native Fuse client



► M Gila (CSCS)



2 Data Namespace: /cvmfs/*.osgstorage.org

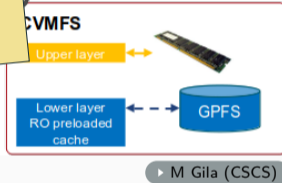




① HPC Client Deployment

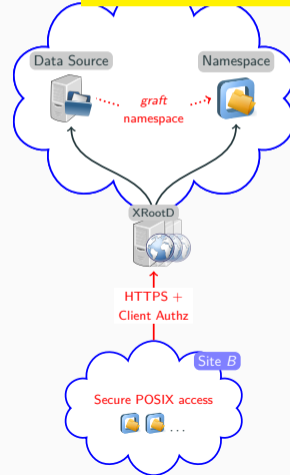
Multi-layer cache +
memory cache plugin

- Runs ATLAS, CMS, and LHCb jobs from native Fuse client



② Data Namespace: `/cvmfs/*.osgstorage.org`

access to multiple PB of data

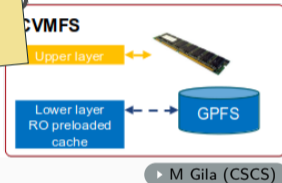




1 HPC Client Deployment

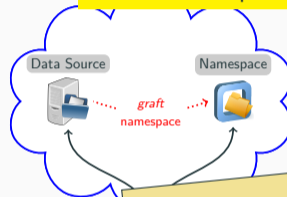
Multi-layer cache +
memory cache plugin

- Runs ATLAS, CMS, and LHCb jobs from native Fuse client



2 Data Namespace: /cvmfs/*.osgstorage.org

access to multiple PB of data



Authentication plugin +
high-bandwidth CDN

Client Authz



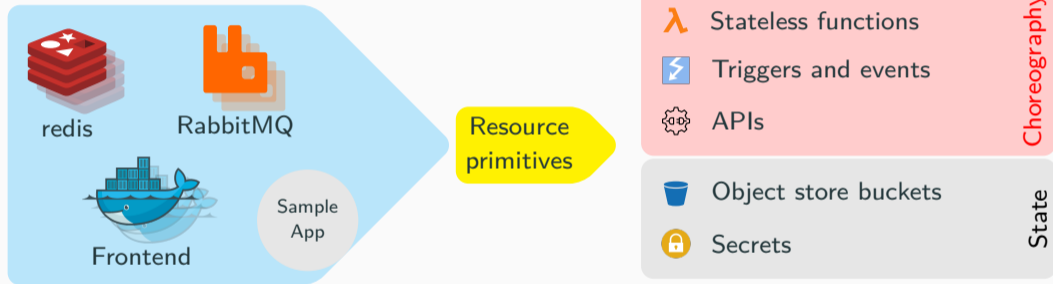
Serverless Publishing

“Serverless” – Looking inside the Box



A reactive model of cloud computing

Replace a set of always-on components by **triggered functions** and **explicit state**



Commercial offers from the big cloud platforms.

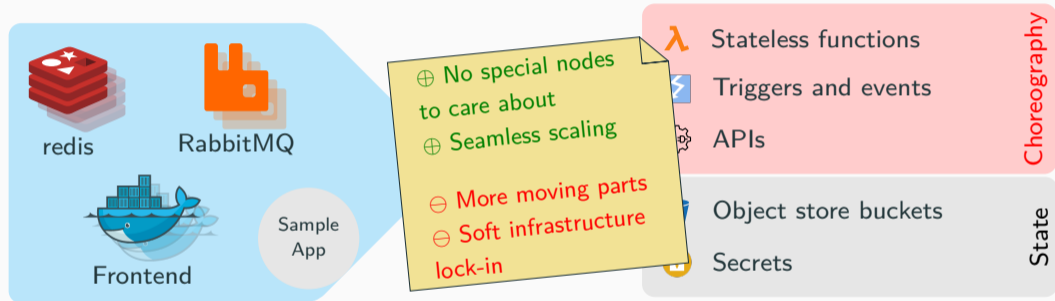
Many open source frameworks emerging, core functionality spawns **ephemeral containers on demand**

“Serverless” – Looking inside the Box



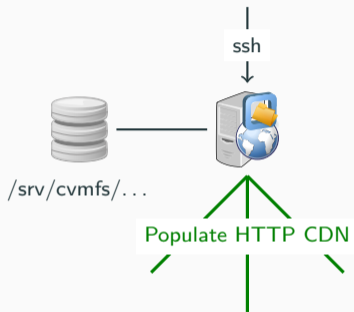
A reactive model of cloud computing

Replace a set of always-on components by **triggered functions** and **explicit state**



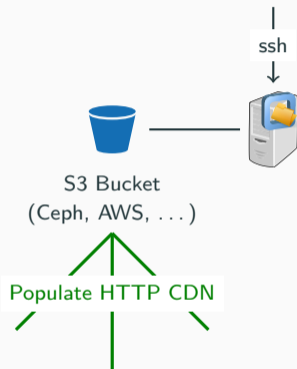
Commercial offers from the big cloud platforms.

Many open source frameworks emerging, core functionality spawns **ephemeral containers on demand**



1 Stand-Alone Release Manager Machine

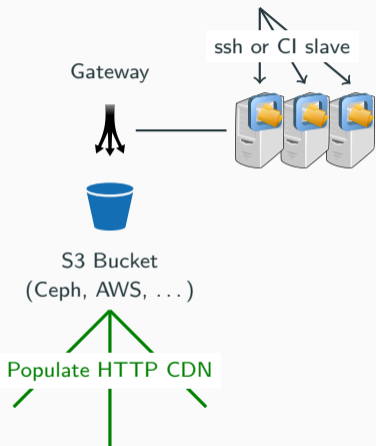
- Dedicated web server
- Local storage or attached block volume



S3 Storage

CHEP'15

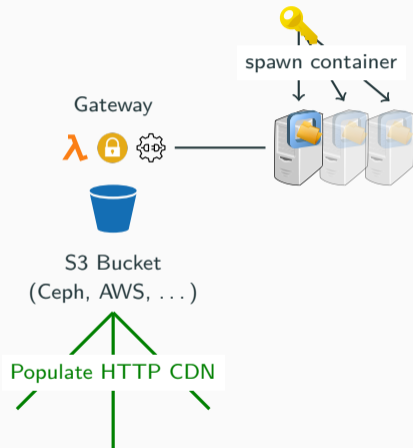
- Dedicated release manager machine, now stateless
- Cloud storage, also available to replicas



③ Multiple Release Manager Machines

Released 2018

- Concurrent release manager access through new gateway services
- Example: `/cvmfs/cernvm-prod.cern.ch`
- Gateway services powered by Erlang/OTP
 - **API** for publishing
 - Issues **leases** for sub paths
 - Receives change sets as set of **signed object packs**



4 Future Serverless Architecture

- On demand release manager container

▶ Prototype Report

- Gateway service primitives

State Access keys and active leases

Functions Lease management,
receiving object packs,
committing change sets



1. Scale number of writers (users, computers) from tens to hundreds per repository.
2. Fast-path ("*portal*") to directly push certain payload types.



```
λ cvmfs enter hsf.cvmfs.io /users/joe
λ # Opens a shell with write access
λ cvmfs publish
λ # Back to read-only mode
```

```
λ cvmfs push docker://hsf/software \
  hsf.cvmfs.io /containers
λ cvmfs push myanalysis.tar.gz \
  hsf.cvmfs.io /users/joe
```



Development program for ~ 2 years.



We have a **stable, yet extensible core** which will continue receiving our attention.

Targeted Out of the Box Support

Read access to /cvmfs on **grids, clouds, supercomputers, end-user devices**
for **production software, integration builds, container images, auxiliary data**

Future Directions

Focus on **scaling up publishing workloads** from tens to hundreds of users per repository

```
λ touch /cvmfs
```

If you'd like to get involved, please contact us!



We have a **stable, yet extensible core** which will continue receiving our attention.

Targeted Out of the Box Support

Read access to /cvmfs on **grids, clouds, supercomputers, end-user devices**
for **production software, integration builds, container images, auxiliary data**

Future Directions

Focus on **scaling up publishing workloads** from tens to hundreds of users per repository

```
λ touch /cvmfs
```

If you'd like to get involved, please contact us!

Thank you for your time!

Backup Slides
