CernVM-FS: Status and Plans

Jakob Blomer (CERN)

CernVM Workshop 2022

Amsterdam, 12 September 2022

Agenda

State of Affairs

New Developments

Container Support

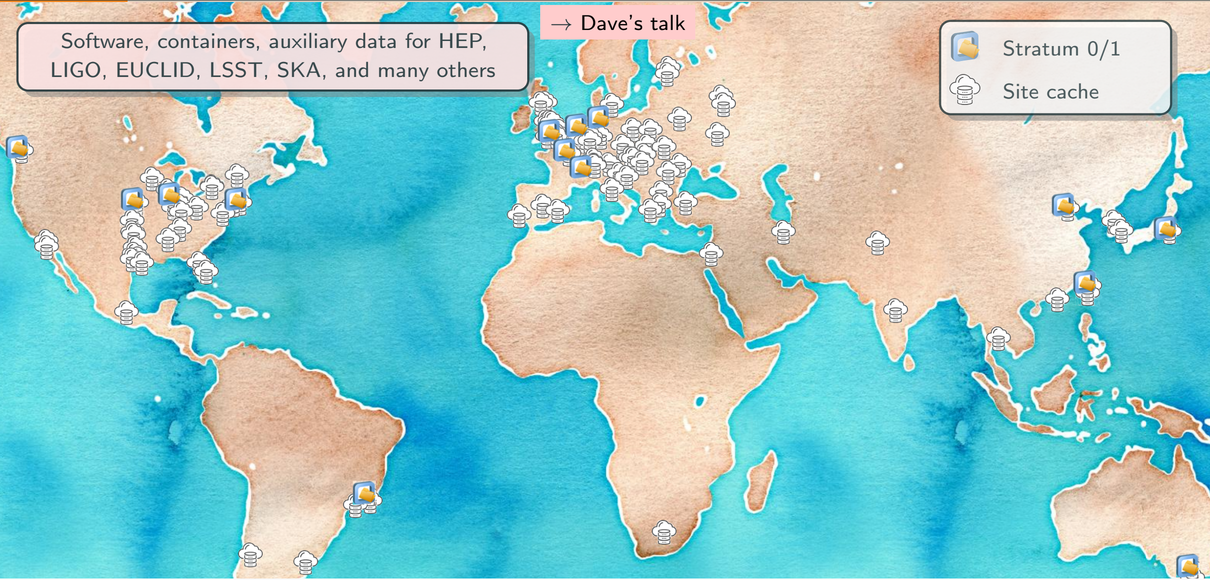
Outlook and Plans

State of Affairs

At a Glance: CernVM-FS Deployment (Grid)

Software, containers, auxiliary data for HEP, LIGO, EUCLID, LSST, SKA, and many others

→ Dave's talk



At a Glance: CernVM-FS Deployment (Grid)

Software, containers, auxiliary data for HEP, LIGO, EUCLID, LSST, SKA, and many others

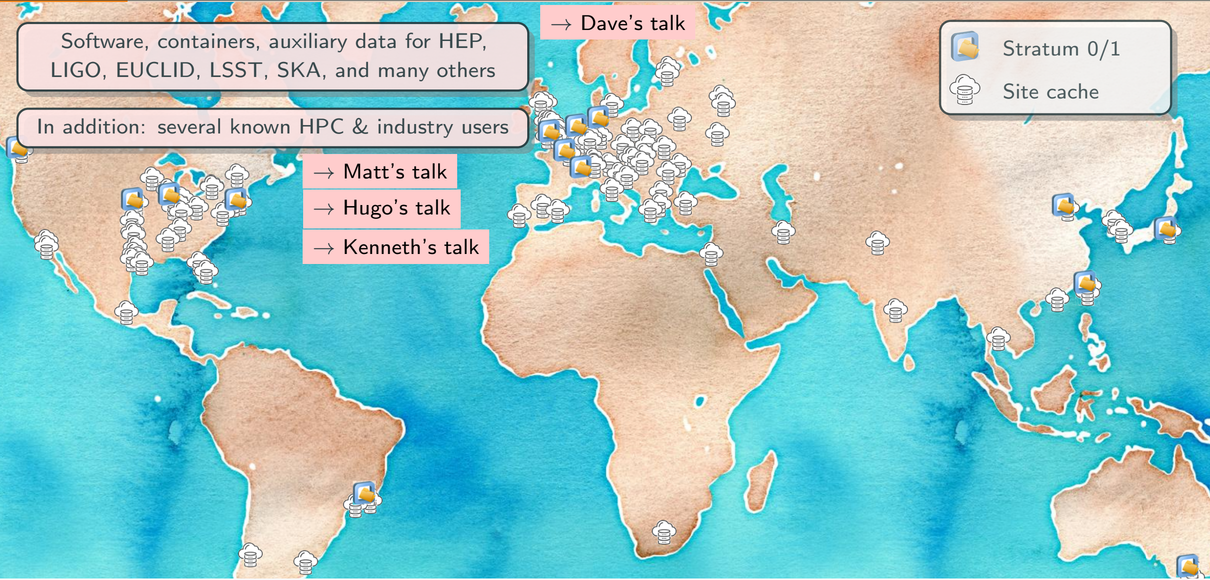
In addition: several known HPC & industry users

→ Matt's talk

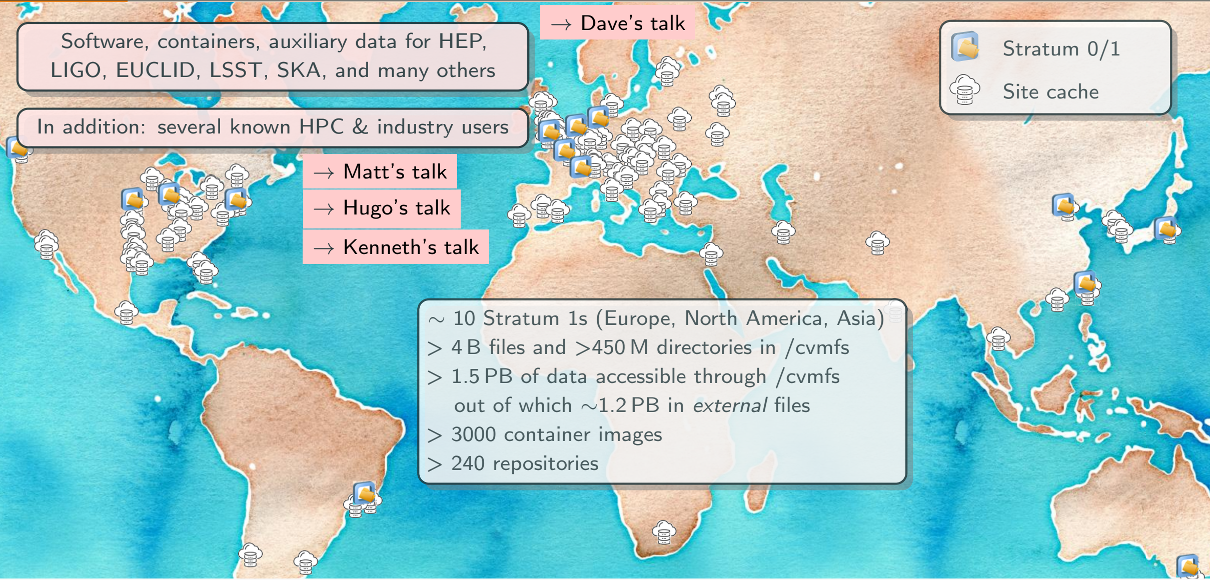
→ Hugo's talk

→ Kenneth's talk

→ Dave's talk



At a Glance: CernVM-FS Deployment (Grid)



Software, containers, auxiliary data for HEP, LIGO, EUCLID, LSST, SKA, and many others

In addition: several known HPC & industry users

→ Dave's talk

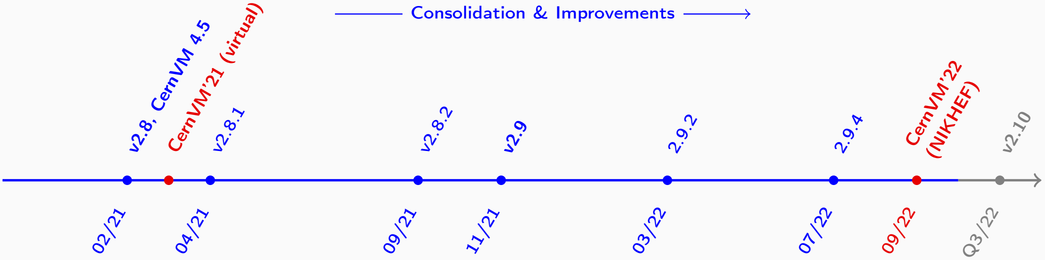
→ Matt's talk

→ Hugo's talk

→ Kenneth's talk

- ~ 10 Stratum 1s (Europe, North America, Asia)
- > 4 B files and >450 M directories in /cvmfs
- > 1.5 PB of data accessible through /cvmfs out of which ~1.2 PB in *external* files
- > 3000 container images
- > 240 repositories

Releases and Events



→ Version 2.10 preview

Highlights of the 2.9 and 2.10 releases

- Performance optimizations in the fuse client and in the S3 & gateway publishers
- Support for proxy sharding
- Support for container registry proxies
- Support for publishing from the ephemeral shell (experimental)

Platform Support

	EL 7	EL 8	EL 9†	Ubuntu 16.04, 18.04	Ubuntu 20.04	Ubuntu 22.04†	Debian 8–10	Debian 11
x86_64	✔	✔	new	✔	✔	new	✔	new
AArch64	✔	new	new	—	new	new		
i686	—	—	—	✔	—	—		

	SLES 12	SLES 15	macOS 11–12‡	Container	WSL 2
x86_64	✔	new	✔	✔	✔
AArch64			✔		—

New platforms added as needed and as build and test hardware is available

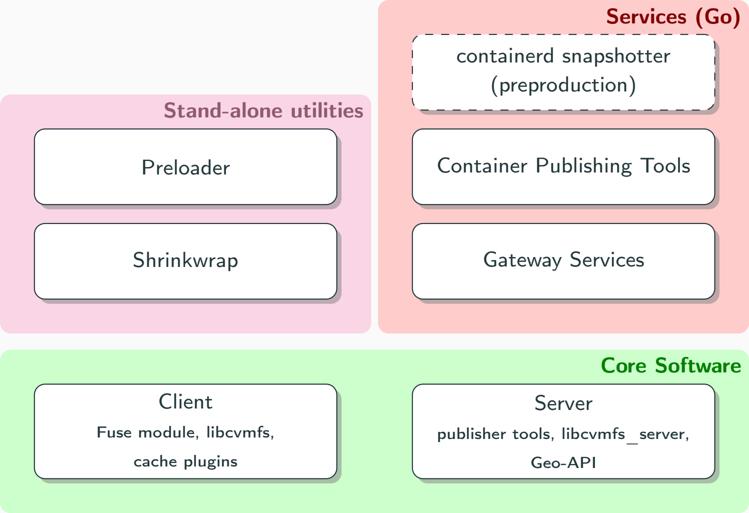
† Required code restructuring for OpenSSL 3

‡ Currently requires osxfuse 3rd party kernel extension – ▸ fuse-t looks like an interesting alternative.
Apple silicon support through Rosetta, native builds still in the roadmap

CernVM-FS Components

Extras:

- cvmfsexec
- cvmfs-servermon
- github-action-cvmfs
- cvmfs-x509-helper
- repository monitor
- . . .



Services (Go)

containerd snapshotter (preproduction)

Container Publishing Tools

Gateway Services

Stand-alone utilities

Preloader

Shrinkwrap

Core Software

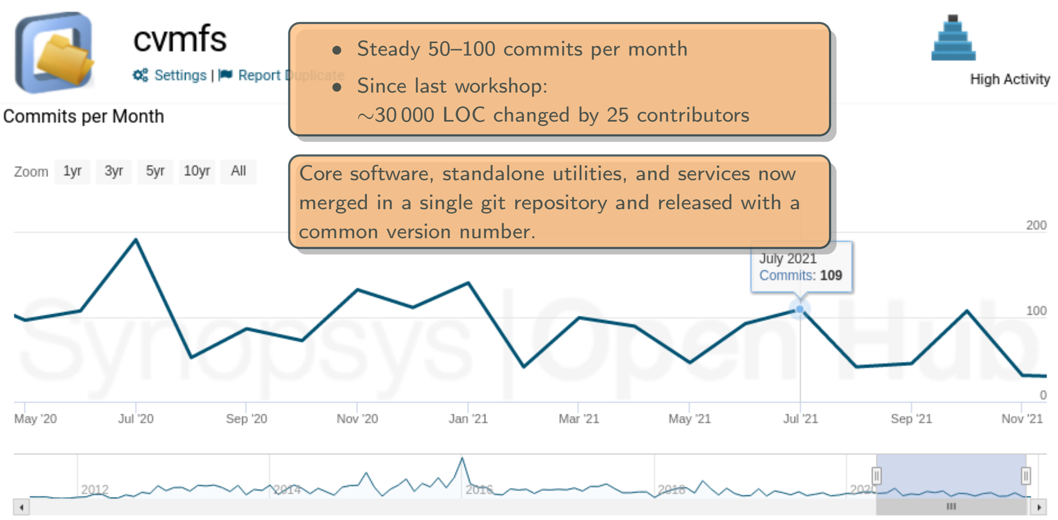
Client

Fuse module, libcvmfs, cache plugins

Server

publisher tools, libcvmfs_server, Geo-API

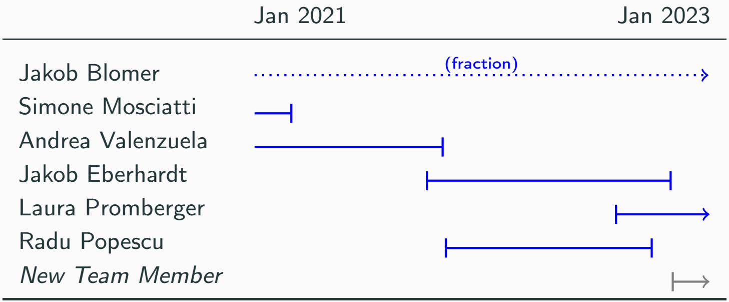
At a Glance: Code



- Steady 50–100 commits per month
- Since last workshop:
  ∼30 000 LOC changed by 25 contributors

Core software, standalone utilities, and services now merged in a single git repository and released with a common version number.

Developer Team @ CERN



Jan 2021

Jan 2023

Jakob Blomer

(fraction)

Simone Mosciatti

Andrea Valenzuela

Jakob Eberhardt

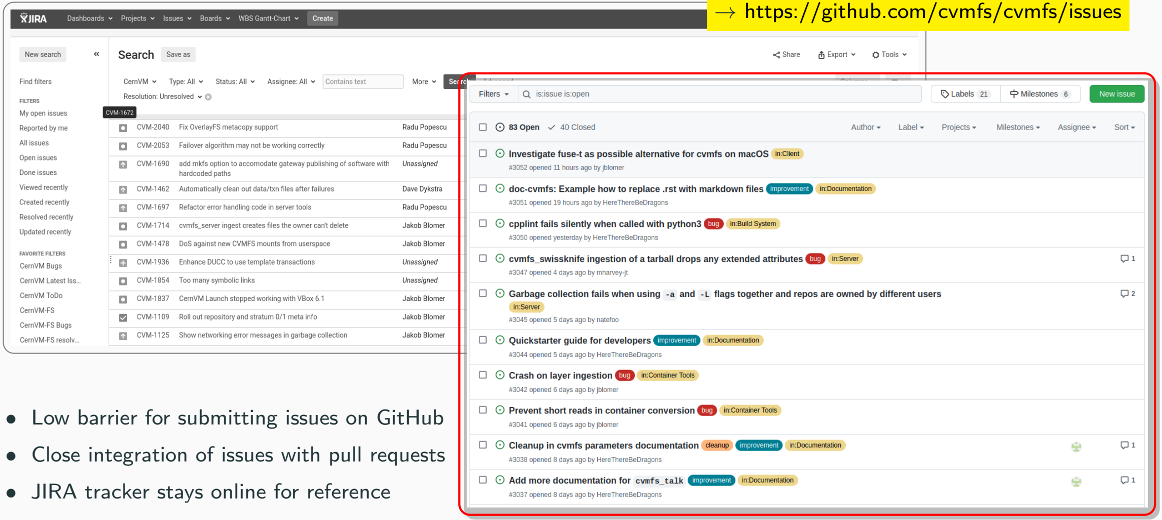
Laura Promberger

Radu Popescu

New Team Member

- Unforeseen departure of Radu to industry; new team member expected by the end of the year
- Laura started a 3 years contract – **huge thanks to Jump Trading** for making that possible!

Issue Tracking: Moved to GitHub



→ https://github.com/cvmfs/cvmfs/issues

Find filters

FILTERS

- My open issues
- Reported by me
- All issues
- Open issues
- Done issues
- Viewed recently
- Created recently
- Resolved recently
- Updated recently

FAVORITE FILTERS

- CernVM Bugs
- CernVM Latest Iss...
- CernVM ToDo
- CernVM-FS
- CernVM-FS Bugs
- CernVM-FS resolv...

CernVM | Type: All | Status: All | Assignee: All | Contains text | More

Resolution: Unresolved

CVM-1672

CVM-2040	Fix OverlayFS metacopy support	Radu Popescu
CVM-2053	Failover algorithm may not be working correctly	Radu Popescu
CVM-1690	add mkfs option to accomodate gateway publishing of software with hardcoded paths	*Unassigned*
CVM-1462	Automatically clean out data/txn files after failures	Dave Dykstra
CVM-1697	Refactor error handling code in server tools	Radu Popescu
CVM-1714	cvmfs_server ingest creates files the owner can't delete	Jakob Blomer
CVM-1478	DoS against new CVMFS mounts from userspace	Jakob Blomer
CVM-1936	Enhance DUCC to use template transactions	*Unassigned*
CVM-1854	Too many symbolic links	*Unassigned*
CVM-1837	CernVM Launch stopped working with VBox 6.1	Jakob Blomer
CVM-1109	Roll out repository and stratum 0/1 meta info	Jakob Blomer
CVM-1125	Show networking error messages in garbage collection	Jakob Blomer

83 Open ✓ 40 Closed

- **Investigate fuse-t as possible alternative for cvmfs on macOS** in:Client
  #3052 opened 11 hours ago by jblomer
- **doc-cvmfs: Example how to replace .rst with markdown files** improvement in:Documentation
  #3051 opened 19 hours ago by HereThereBeDragons
- **cpplint fails silently when called with python3** bug in:Build System
  #3050 opened yesterday by HereThereBeDragons
- **cvmfs_swissknife ingestion of a tarball drops any extended attributes** bug in:Server — 1
  #3047 opened 4 days ago by mharvey-jt
- **Garbage collection fails when using `-a` and `-L` flags together and repos are owned by different users** in:Server — 2
  #3045 opened 5 days ago by natefoo
- **Quickstarter guide for developers** improvement in:Documentation
  #3044 opened 5 days ago by HereThereBeDragons
- **Crash on layer ingestion** bug in:Container Tools
  #3042 opened 6 days ago by jblomer
- **Prevent short reads in container conversion** bug in:Container Tools
  #3041 opened 6 days ago by jblomer
- **Cleanup in cvmfs parameters documentation** cleanup improvement in:Documentation — 1
  #3038 opened 8 days ago by HereThereBeDragons
- **Add more documentation for `cvmfs_talk`** improvement in:Documentation — 1
  #3037 opened 8 days ago by HereThereBeDragons

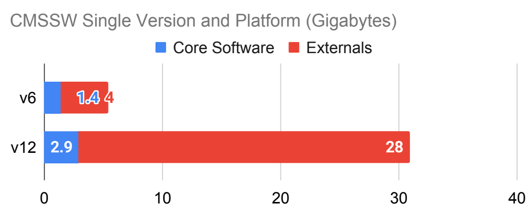
- Low barrier for submitting issues on GitHub
- Close integration of issues with pull requests
- JIRA tracker stays online for reference
- Updating existing tickets still possible

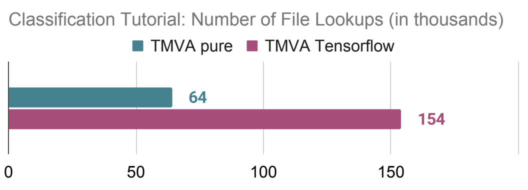
New Developments

Software Management for HL-LHC

Compared to LHC Run 1-2 (2011–2018), we now find

- Multiple target architectures: AArch64, x86_64 micro-architectures (e.g. AVX512), IBM Power, GPUs
- A growing Python software ecosystem, in particular for machine learning tasks
- More agile software development: automated integration builds, nightly builds
- Many more cores per box
- Deployment with containers


CMSSW Single Version and Platform (Gigabytes)


Classification Tutorial: Number of File Lookups (in thousands)

My estimate: the software distribution problem for HL-LHC grows by a factor of 3-5 for most key metrics.

→ We should invest in the CernVM-FS performance, scalability, and correctness of edge cases

Improved Page Cache Management in the Fuse Client

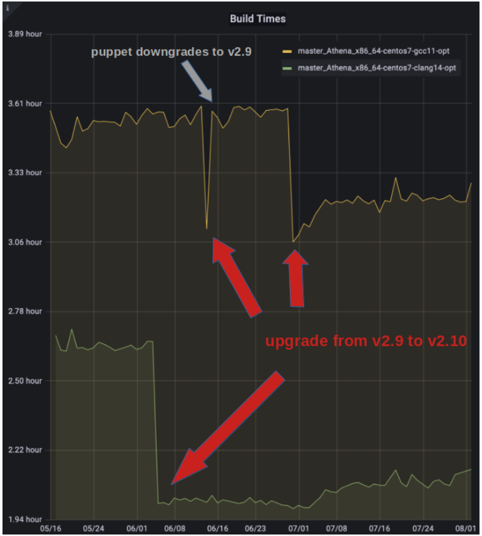
Problem

Bad CPU utilization when building ATLAS Athena on 64+ core nodes; compiler loaded from CernVM-FS.

Caused by very limited used of kernel page cache for data by the fuse client <2.10.

Key issue addressed in version 2.10 is purging of the caches when the file content changes.

→ 10 % to 30 % faster Athena builds

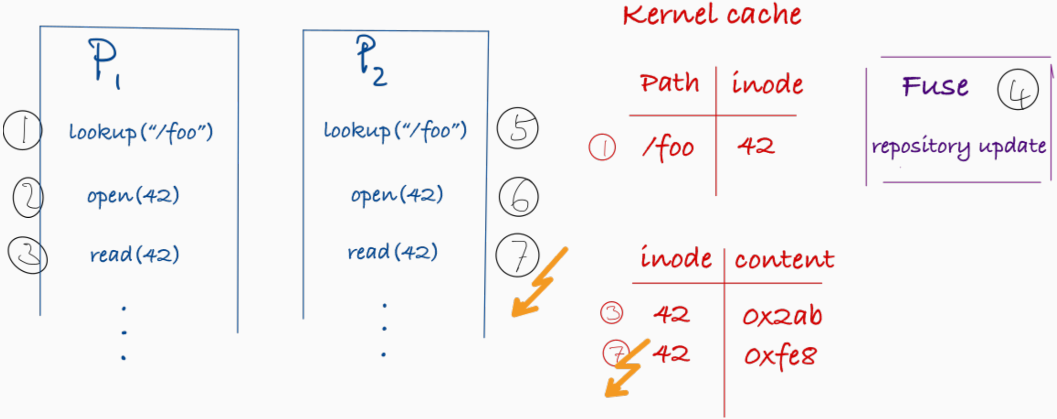


Johannes Elmsheuser (ATLAS)

Fixed Handling of Open, Changing Files

Fixed in version 2.10

A file that is concurrently read in two different version can return corrupted content – surprisingly only recently triggered by Compute Canada and EESSI



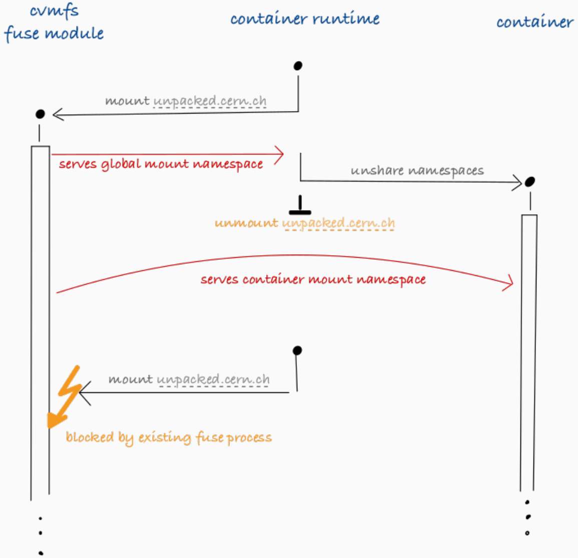
	Path	inode
①	/foo	42

	inode	content
③	42	0x2ab
⑦	42	0xfe8

Fixed Zombie Mountpoints

Fixed in version 2.9 + Kernel 5.15 (EL 9.1)

- Depending on the container engine (use of `unshare`), mounting a repository could hang
- Fixed by allowing new mounts to attach to existing fuse module
- Got us a mention on ▸ phoronix



Next Steps

Evaluate proxy sharding

available in version 2.10

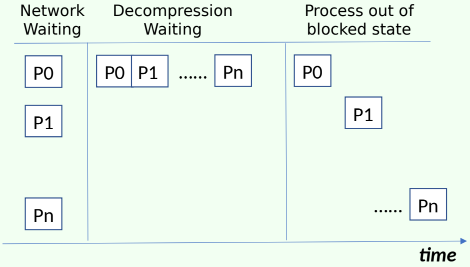
- Should reduce cold cache latency with multiple proxies

Kernel caching of symlink resolution

- Challenge: preserve cache consistency across file system updates
- Requires patching ▸ fuse, currently being tested

Improve cold cache performance on many-core nodes

- Concurrent download streams are stalled by serialized decompression



Razvan Virtan

Container Support

CernVM-FS as a Container Hub

/cvmfs/unpacked.cern.ch

- > 2200 images
- > 10 TB
- > 250 M files

/cvmfs/singularity.opensciencegrid.org

- > 900 images
- > 3.5 TB
- > 75 M files

Images are readily available to run with apptainer (singularity), including **base operating systems**, **experiment software stacks**, **explorative tools (ML etc.)**, **user analyses**, and special-purpose containers such as **folding@home**

```
$ /cvmfs/oasis.opensciencegrid.org/mis/apptainer/current/bin/apptainer \
  exec '/cvmfs/unpacked.cern.ch/registry.hub.docker.com/library/debian:stable' \
  cat /etc/issue
Debian GNU/Linux 11 \n \l
```

CernVM-FS as a Container Hub

/cvmfs/unpacked.cern.ch

- > 2200 images
- > 10 TB
- > 250 M files

/cvmfs/singularity.opensciencegrid.org

- > 900 images
- > 3.5 TB
- > 75 M files

> 2× growth since 2021 workshop

Images are readily available to run with apptainer (singularity), including **base operating systems**, **experiment software stacks**, **explorative tools (ML etc.)**, **user analyses**, and special-purpose containers such as **folding@home**

```
$ /cvmfs/oasis.opensciencegrid.org/mis/apptainer/current/bin/apptainer \
  exec '/cvmfs/unpacked.cern.ch/registry.hub.docker.com/library/debian:stable' \
  cat /etc/issue
Debian GNU/Linux 11 \n \l
```

Container Runtime Integration

Runtime	CernVM-FS Support
Apptainer	native
podman	native / pre-production (use image storage from /cvmfs)
containerd / k8s	plugin / pre-production (through cvmfs snapshotter)
docker	*"graph driver"* image storage plugin – deprecated[1]
	through containerd in the future

→ Kohei's talk

Documentation chapter on containers & CernVM-FS:

→ `https://cvmfs.readthedocs.io/en/latest/cpt-containers.html`

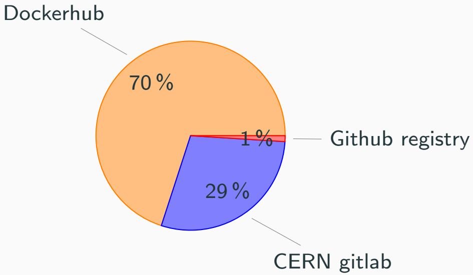
[1] Soon replaced by containerd ▸ Docker's announcement

unpacked.cern.ch

- Image wishlists on ▸ CERN GitLab and ▸ GitHub
- Editable by merge/pull request

```
version: 1
user: cvmfsunpacker
cvmfs_repo: 'unpacked.cern.ch'
output_format: >
  https://gitlab-registry.cern.ch/unpacked/sync/$(image)
input:
   - 'https://gitlab-registry.cern.ch/sft/docker/ubuntu20:latest'
   - 'https://registry.hub.docker.com/library/centos:*'
   ...
```

Origin of images on unpacked.cern.ch



Next steps:

- Complete podman store
- Multi-arch image support
- Release webhook integration with Harbor

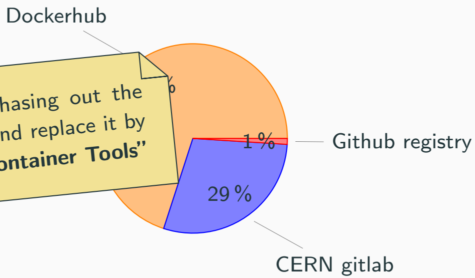
new Images from Docker Hub and GitHub are proxied through registry.cern.ch → Ricardo's talk

unpacked.cern.ch

- Image wishlists on CERN GitLab and GitHub
- Editable by merge/pull request

```
version: 1
user: cvmfsunpacker
cvmfs_repo: 'unpacked.cern.ch'
output_format: >
  https://gitlab-registry.cern.ch/unpacked/sync/$(im
input:
   - 'https://gitlab-registry.cern.ch/sft/docker/ubu
   - 'https://registry.hub.docker.com/library/centos:
   ...
```

Origin of images on unpacked.cern.ch



Note: we are phasing out the name "DUCC" and replace it by **"CernVM-FS Container Tools"**

Next steps:

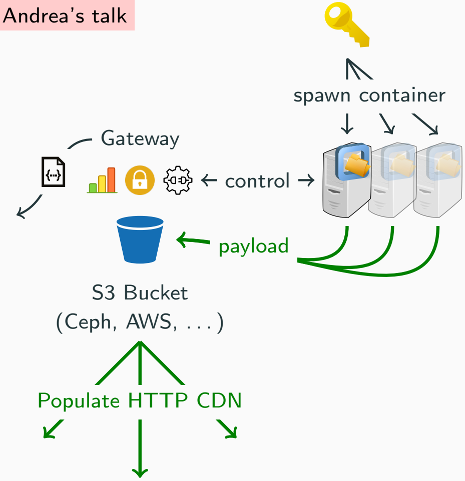
- Complete podman store
- Multi-arch image support
- Release webhook integration with Harbor

new Images from Docker Hub and GitHub are proxied through registry.cern.ch → Ricardo's talk

Outlook and Plans

Progress towards Containerized Publishing

→ Andrea's talk



Goal for Final Setup

- On-demand publish container
- Gateway services:
  - Provides **API** for publishing
  - Issues **leases** for sub paths
  - Updates **repository statistics**
- All components deployable on k8s

Component Status

S3 backend	**production**
Gateway service	**robust**, with known issues
Publish container	**prototype**

Links

Source code	`https://github.com/cvmfs/cvmfs`
Documentation	`https://cvmfs.readthedocs.io`
Support forum	`https://cernvm-forum.cern.ch`
Mattermost	`https://mattermost.web.cern.ch/cernvm`
Bug tracker	`https://github.com/cvmfs/cvmfs/issues` new
Package repositories	`https://cvmrepo.s3.cern.ch/` new

Summary & Next Milestones

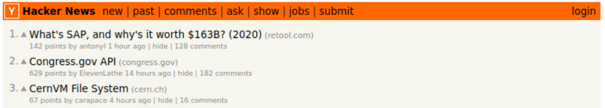
Goal: prepare CernVM-FS for software distribution at HL-LHC

1. Continued client-side performance engineering
2. Two main publisher workflows
   - guarded by software & dataset librarians
   - container ingestion open to a broader community
3. Address missing functionality in the gateway to make it work together with the container tools
4. Container integration with containerd/k8s and podman:
   releases of pre-production code, documentation, packaging (e.g. helm charts)
5. Balance new developments with maintenance (platforms, code infrastructure, . . .)

Summary & Next Milestones

Goal: prepare CernVM-FS for software distribution at HL-LHC

1. Continued client-side performance engineering
2. Two main publisher workflows
   - guarded by software & dataset librarians
   - container ingestion open to a broader community
3. Address missing functionality in the gateway to make it work together with the container tools
4. Container integration with containerd/k8s and podman:
   releases of pre-production code, documentation, packaging (e.g. helm charts)
5. Balance new developments with maintenance (platforms, code infrastructure, . . .)


Y
Hacker News new | past | comments | ask | show | jobs | submit
login

1. ▲ What's SAP, and why's it worth $163B? (2020) (retool.com)
142 points by antonyl 1 hour ago | hide | 128 comments

2. ▲ Congress.gov API (congress.gov)
629 points by ElevenLathe 14 hours ago | hide | 182 comments

3. ▲ CernVM File System (cern.ch)
67 points by carapace 4 hours ago | hide | 16 comments

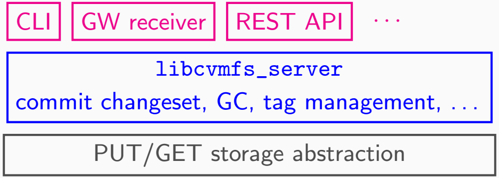
Backup Slides

Next-Generation Server Code

Legacy Code

A set of tools targeted for a dedicated release manager machine, and the interactive workflow *open transaction* + *copy* + *commit*

New Architecture



CLI

GW receiver

REST API

· · ·

`libcvmfs_server`

commit changeset, GC, tag management, . . .

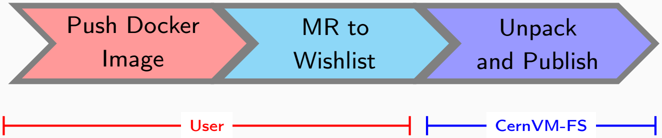
PUT/GET storage abstraction

A common base library providing repository transformation primitives, on top of which higher-level publish abstractions can be built

Initial CLI commands ported to `libcvmfs_server`: `info`, `diff`, `transaction`, `enter`.

Foundation for new functionality and workflows, e.g. template transactions, ephemeral writable shell

Container Conversion Service



Wishlist https://gitlab.cern.ch/unpacked/sync

```yaml
version: 1
user: cvmfsunpacker
cvmfs_repo: 'unpacked.cern.ch'
output_format: >
  https://gitlab-registry.cern.ch/unpacked/sync/$(image)
input:
    - 'https://registry.hub.docker.com/library/fedora:latest'
    - 'https://registry.hub.docker.com/library/debian:stable'
    - 'https://registry.hub.docker.com/library/centos:*'
```

Multiple wishlists possible, e.g. experiment specific

/cvmfs/unpacked.cern.ch

```
# Singularity
/registry.hub.docker.com/fedora:latest -> \
  /cvmfs/unpacked.cern.ch/.flat/d0/d0932...
# containerd, k8s, podman
/.layers/f0/1af7...
```

Simple Case: CernVM-FS Available on the Host

```
$ docker run -v /cvmfs:/cvmfs:shared busybox ls /cvmfs/sft.cern.ch
README.md lcg
```

```
$ singularity exec -B /cvmfs docker://busybox ls /cvmfs/sft.cern.ch
README.md lcg
```

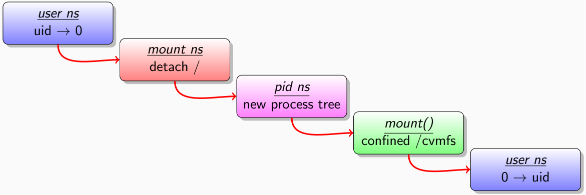
Important: use *shared* bind mount with docker so that that repositories can be mounted on demand from inside the container

Unprivileged Mounting with `cvmfsexec`

```
$ cvmfsexec grid.cern.ch atlas.cern.ch -- ls /cvmfs
atlas.cern.ch cvmfs-config.cern.ch grid.cern.ch
```

Technical foundations

- User namespaces completing container support
- As of Linux kernel version 4.18 (EL8, but also EL 7.8), **fuse mounts are unprivileged in user name spaces**
- Overlay-FS implementation available as a fuse module



user ns
uid → 0

mount ns
detach /

pid ns
new process tree

mount()
confined /cvmfs

user ns
0 → uid

For HPCs: Pre-mounted by Singularity

- With the new Fuse3 libraries, mounting can be handed off to a trusted, external helper.
- Fuse3 libraries have been backported to EL6 and EL7 platforms.
- Gives access to /cvmfs in containers started by singularity (`singularity --fusemount`)
- **Required cvmfs client to be installed and prepared in the container**

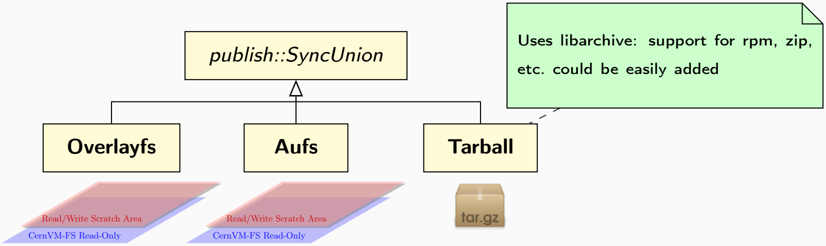
```
$ CONFIGREPO=config-osg.opensciencegrid.org
$ mkdir -p $HOME/cvmfs_cache
$ singularity exec -S /var/run/cvmfs -B $HOME/cvmfs_cache:/var/lib/cvmfs \
  --fusemount "container:cvmfs2 $CONFIGREPO /cvmfs/$CONFIGREPO" \
  --fusemount "container:cvmfs2 sft.cern.ch /cvmfs/sft.cern.ch" \
  docker://davedykstra/cvmfs-fuse3 ls /cvmfs/sft.cern.ch
README.md lcg
```

Enabling Feature for Container Publishing: Tarball Ingestion

Direct path for the common pattern of publishing tarball contents

```
$ cvmfs_server transaction
$ tar -xf ubuntu.tar.gz
$ cvmfs_server publish
```

```
$ cat ubuntu.tar.gz | \
    cvmfs_server ingest -t -
```



publish::SyncUnion

Overlayfs

Aufs

Tarball

Read/Write Scratch Area

CernVM-FS Read-Only

Read/Write Scratch Area

CernVM-FS Read-Only

tar.gz

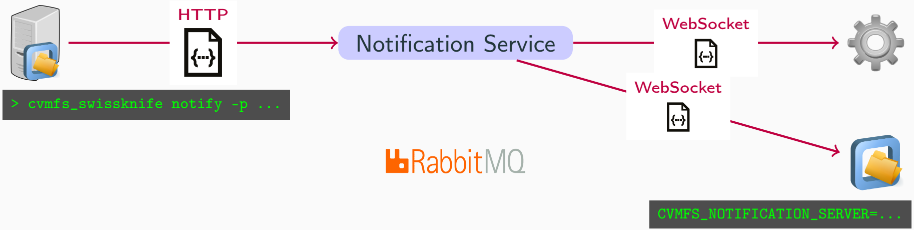
Uses libarchive: support for rpm, zip, etc. could be easily added

Performance Example

Ubuntu 18.04 container – 4 GB in 250 k files: **56 s untar + 1 min publish vs. 74s ingest**

Notification Service

Fast distribution channel for repository manifest: useful for CI pipelines, data QA



```
> cvmfs_swissknife notify -p ...
```

```
CVMFS_NOTIFICATION_SERVER=...
```

- Optional service supporting a regular repository
- Publish/subscribe utility in `cvmfs_swissknife`
- Subscribe component integrated with the client, automatic reload on changes

→ **CernVM-FS writing remains asynchronous but with fast response time in $\mathcal{O}$(seconds)**